Head First Statistics

深入浅出统计学



轻松剔除
图形错误



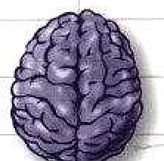
利用标准差提高
赛季平均成绩





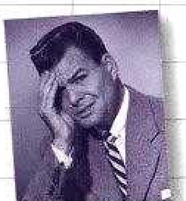
赌场趋吉避凶

让统计概念
植根于大脑





看统计学如何
瞒天过海



避免难堪的
抽样失误

Dawn Griffiths 著
李芳 译

O'REILLY®

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

深入浅出统计学

统计/数学

你将从本书学到什么？

如果有一本探讨统计学的书籍，能够让直方图（histogram）、概率分布（probability distribution）、卡方分析（chi square analysis）的学习不再像看牙医那么恐怖，那该有多好？正是《深入浅出统计学》这一本书，为这个枯燥的领域带来鲜活的乐趣，运用充满互动性的真实世界情节，教导你有关这门学科的所有基础，内容涵盖甚广，从分析运动比赛、博弈游戏到临床药物试验都有。

不管你是在修习统计学，准备统计学考试，或者只是对统计分析抱着极大的好奇心，“深入浅出”系列的撰写风格都能为你提供莫大的帮助，不仅让你充分掌握统计学的要义，更会告诉你如何将统计理论应用到日常生活中。



本书为何与众不同？

我们认为你的时间极其宝贵，不该浪费在冥思苦想各种新名词、新概念上。《深入浅出统计学》运用认知科学与学习理论的最新研究成果，精心建构出一段引发多重感知的学习体验。《深入浅出统计学》采取专为大脑运作而设计的丰富视觉化风格，你将不再被密密麻麻的文字催得昏昏欲睡。

“《深入浅出统计学》是目前市面上最具娱乐性、最能够抓住读者注意力的统计学研读指南。透过生动活泼的手法与素材，为这个困难的主题提供最容易被接受的学习方式，贯穿全书的精辟解说让各种程度的学生都能够充分地理解统计学的妙义。”

——阿瑞娜·安德森（Ariana Anderson），加利福尼亚大学洛杉矶分校统计系教师助理及博士生

“《深入浅出统计学》运用简单的生活实例，提供最符合直觉的理解方式，让统计理论的学习既有趣又自然。”

——迈克尔·普瑞诺（Michael Prerau），波士顿大学计算神经科学和统计学讲师

图书分类：统计/数学

责任编辑：李影

www.phei.com.cn

本书贴有激光防伪标志，凡没有防伪标志者，属盗版图书。

O'REILLY®

oreilly.com.cn
headfirstlabs.com

O'Reilly Media, Inc.授权电子工业出版社出版

此简体中文版仅限于中国大陆（不包含中国香港、澳门特别行政区和中国台湾地区）销售发行
This Authorized Edition for sale in the mainland of China (excluding Hong Kong, Macao and Taiwan)

ISBN 978-7-121-15308-2

定价：89.00元

O'REILLY®

深入浅出统计学

Head First Statistics



Dawn Griffiths 著

李芳 译

電子工業出版社

Publishing House of Electronics Industry

北京・BEIJING

内容简介

《深入浅出统计学》具有深入浅出系列的一贯特色，提供最符合直觉的理解方式，让统计理论的学习既有趣又自然。从应对考试到解决实际问题，无论你是学生还是数据分析师，都能从中受益。本书涵盖的知识点包括：信息可视化、概率计算、几何分布、二项分布及泊松分布、正态分布、统计抽样、置信区间的构建、假设检验、卡方分布、相关与回归等等，完整涵盖AP考试范围。本书运用充满互动性的真实世界情节，教给你有关这门学科的所有基础，为这个枯燥的领域带来鲜活的乐趣，不仅让你充分掌握统计学的要义，更会告诉你如何将统计理论应用到日常生活中。

978-0-596-52758-7 Head First Statistics © 2009 by O'Reilly Media, Inc. Simplified Chinese edition, jointly published by O'Reilly Media ,Inc. and Publishing House of Electronics Industry, 2011.Authorized translation of the English edition, 2009 O'Reilly Media, Inc., the owner of all rights to publish and sell the same.All rights reserved including the rights of reproduction in whole or in part in any form.

本书中文简体版专有出版权由 O'Reilly Media, Inc. 授予电子工业出版社，未经许可，不得以任何方式复制或抄袭本书的任何部分。

版权贸易合同登记号　图字：01-2011-7144

图书在版编目（CIP）数据

深入浅出统计学 /（美）格里菲思（Griffiths,D.）著；李芳译 . — 北京：电子工业出版社，2012.1 书名原文：Head First Statistics

ISBN 978-7-121-15308-2

Ⅰ.①深… Ⅱ.①格… ②李… Ⅲ.①统计学—通俗读物 Ⅳ.① C8-49

中国版本图书馆 CIP 数据核字 (2011) 第 243859 号

责任编辑：李　影　　　　策　划：卢鹓翔

印　　刷：北京天宇星印刷厂

装　　订：三河市鹏成印业有限公司

出版发行：电子工业出版社

　　　　　北京市海淀区万寿路 173 信箱　　　邮　编：100036

开　　本：860×1092　1/16　　印　张：45　　字　数：717 千字

印　　次：2012 年 1 月第 1 次印刷

印　　数：4 000 册　　　　定　价：89.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。服务热线：（010）88258888。

“《深入浅出统计学》是目前市面上最具娱乐性、最能够抓住读者注意力的统计学研读指南。透过生动活泼的手法与素材，为这个困难的主题提供最容易被接受的学习方式，贯穿全书的精辟解说让各种程度的学生都能够充分地理解统计学的妙义。”

—— 阿瑞娜·安德森（Ariana Anderson），加利福尼亚大学洛杉矶分校统计系教师助理及博士生

“《深入浅出统计学》润物细无声。当一口气看完讲解和练习后，你就会发现自己在社交谈话中可以开口闭口正态分布、泊松分布，我保证并没有人建议你这么做！”

—— 加里·沃尔夫（Gary Wolf），《连线》杂志（Wired Magazine）特约编辑

“道恩·格里菲思把一些十分复杂的概念拆分为一块块小材料，它们不那么令人望而生畏，凡夫俗子都会觉得十分容易掌握。大量图形、图片让材料具体生动，458页那位吵着要买口香糖球的迷人女模特已然让我心生情愫。”

—— 布鲁斯·弗雷（Bruce Frey），《统计学技巧》（Statistics Hacks）作者

“《深入浅出统计学》运用简单的生活实例，提供最符合直觉的理解方式，让统计理论的学习既有趣又自然。”

—— 迈克尔·普瑞诺（Michael Prerau），波士顿大学计算神经科学和统计学讲师

“你以为‘深入浅出’图书只适合计算机迷吗？不妨试试用本书提供的方式学习统计学，你就会改变想法。这方法的确有用。”

—— 安迪·帕克（Andy Parker）

“这本书非常适合学生学习统计学——寓教于乐、讲解全面、易于理解。完美无缺的方法！”

—— 丹妮尔·莱维特（Danielle Levitt）

“打倒其他枯燥无味的统计书！连我的猫都喜欢这一本。”

—— 凯里·科利特（Cary Collett）

深入浅出系列图书美誉

"凯西和伯特合著的《深入浅出Java》让白纸黑字摇身一变，成为读者领略过的最接近图形用户界面的作品。作者以幽默、新潮的风格，让学习Java成为不断追问'他们后来怎么样了？'的愉快体验。"

—— 沃伦·柯费尔（Warren Keuffel），《软件开发》杂志（Software Development）

"《深入浅出Java》引人入胜的风格会把本来一无所知的你变成斗志昂扬的Java战士，不仅如此，书中还收入了大量实用事例，这样的实用事例在其他文章中只会留给恐怖的'读者练习'。此书睿智、幽默、新潮而实用——能在讲授对象序列化和网络加载协议知识的同时有这样的主张并坚持做到的书籍并不多见。

—— 丹·罗素（Dan Russell）博士，IBM Almaden研究中心用户科学和用户体验研究室主任（兼斯坦福大学人工智能教师）

"此书明快，风趣，玩世不恭，引人入胜。细心读——你可能确实能学到东西！"

—— 肯·阿诺德（Ken Arnold），曾任Sun Microsystems高级工程师，与Java创始人詹姆斯·戈斯林（James Gosling）合著《Java编程语言》（The Java Programming Language）

"如醍醐灌顶，脑海中堆积如山的书本知识一下子消化了。"

—— 沃德·坎宁安（Ward Cunningham），维基概念发明人，Hillside Group创立人

"正合我们这些喜欢研究技术、生活随意的程序员的口味，实用开发策略的称手参考书——让我的大脑尽情运转，无须硬着头皮应付迂腐乏味的专家说教。"

—— 特拉维斯·卡兰尼克（Travis Kalanick），Scour网站和Red Swoosh网站创始人，获麻省理工学院TR100（《技术回顾》世界百名青年创新学者）称号

"有的书是用来买的，有的书是用来藏的，还有的书是用来摆在案头的。感谢O'Reilly和Head First的员工，他们出了最高等级的书——深入浅出系列，让人爱不释手、百读不厌。《深入浅出SQL》是我最心爱的书，连参考用的PDF打印稿都快翻烂了。"

—— 比尔·索耶（Bill Sawyer），Oracle公司ATG课程经理

“本书的透彻、幽默和睿智令人钦佩，连编程门外汉也能借助这样的书想出办法解决问题。”

—— 科利·多克托罗（Cory Doctorow），博客网站BoingBoing撰稿人合作编辑，著有《魔法王国的故事》（Down and Out in the Magic Kingdom）及《人来人往的城市》（Someone Comes to Town, Someone Leaves Town）

“昨天收到书就开始读……一读就停不下来了，真是酷毙了。书很有趣，内容扎实，切中肯綮。印象太好了。”

—— 艾瑞克·伽玛（Erich Gamma），IBM杰出工程师，《设计模式》（Design Patterns）合著者

“我读过的最有趣、最高明的软件设计图书之一。”

—— 艾伦·拉伯奇（Aaron LaBerge），ESPN.com技术副主席

“过去要犯着错误摸索前进的漫长学习过程，现在干净利落地浓缩在一本迷人的平装书中。”

—— 麦克·戴维森（Mike Davidson），Newsvine, Inc.首席执行官

“每一章都凝聚着优雅的设计，每一条原理无不饱含实用价值与闪光智慧。”

—— 肯·戈德斯坦（Ken Goldstein），迪斯尼在线执行副总裁

“我 ♥《深入浅出HTML与CSS & XHTML》。它以‘有趣’的模式，将全部知识倾囊相授。”

—— 丹妮尔·莱维特（Danielle Levitt）

“通常，阅读设计模式方面的书或文章时，我都得头悬梁锥刺股才能保证注意力集中。这本书却是个例外，听起来可能有点怪，这本书让学习设计模式变得盎然有趣。

“当其他设计模式方面的书籍还在教读者呀呀学语时，这本书却已在踏浪高歌‘加油，兄弟！’”

—— 艾维克·伍勒（Eric Wuehler）

“我实实在在爱这本书。事实上，我当着老婆的面亲了这本书。”

—— 塞提斯·库玛（Satish Kumar）

O'Reilly其他相关图书

Statistics Hacks™

Statistics in a Nutshell

Mind Hacks™

Mind Performance Hacks™

Your Brain: The Missing Manual

O'Reilly深入浅出系列其他图书

Head First Java™

Head First Object-Oriented Analysis and Design (OOA&D)

Head First HTML with CSS and XHTML

Head First Design Patterns

Head First Servlets and JSP

Head First EJB

Head First PMP

Head First SQL

Head First Programming (2008)

Head First Design Patterns

Head First Servlets and JSP

Head First EJB

Head First PMP

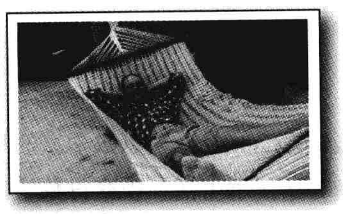
Head First SQL

Head First Programming (2008)

Head First PHP & MySQL (2008)

Head First Algebra (2008)

献给我的爸爸、妈妈、大卫和卡尔。谢谢你们的支持和信任。不过还要等一段时间才能给你们车！

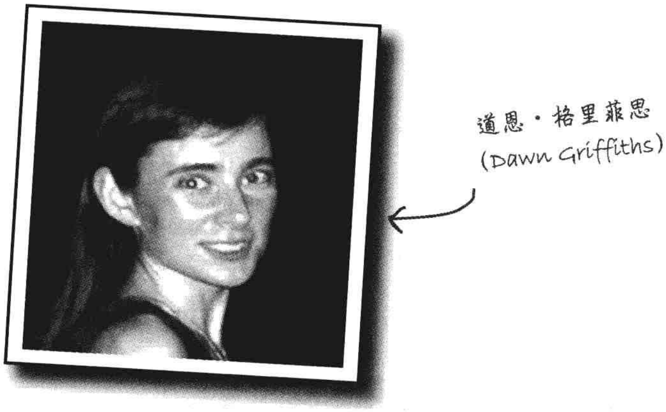


大卫



妈妈和爸爸

作者介绍



道恩·格里菲思（Dawn Griffiths）以数学家的身份在一所顶尖英国大学开始读书生活，曾获得数学专业一级荣誉学位。可是，当她意识到人们在聚会上不愿意与她谈话时，就拒绝了一份研究极为罕见的微分方程的博士奖学金，转而从事软件开发事业。目前她将IT咨询、写作及数学集于一身。

当道恩不写深入浅出系列图书的时候，你会看到她在打太极拳、织梭结花边或在烹饪美食 —— 她还没有修炼出一心三用的绝技。旅行以及和她深爱的丈夫大卫在一起消磨时光也让她十分享受。

道恩有一个猜测 ——《深入浅出梭结花边》有可能会很火爆，可她怀疑布雷特（Brett）和劳里（Laurie）不会同意。

译者序

《深入浅出统计学》翻译好了。这是我翻译的第二本“深入浅出”系列书籍，第一本是《深入浅出数据分析》，两本书可以说是姊妹篇。

《深入浅出统计学》具有深入浅出系列的一贯特色，即“认为深入浅出系列图书的每一位读者都是一位学习者”，遵守“将知识图形化”、“采用对话式的个性化风格”、“引导读者深入思考”、“牢牢吸引读者的注意力”、“影响读者的情感”等约定。不过，这并不意味着无法捕捉到作者的个人特点。

与《深入浅出数据分析》喧嚣错杂的现场气氛相比，《深入浅出统计学》显示出一分属于“办公室”的宁静，由数据分析师整理好的数据就摆在那里，找到合适的方法进行演算成为工作的重点，这些方法包括：信息图形化、集中趋势的量度、分散性与变异性的量度、概率计算、离散概率分布、排列与组合、几何分布、二项分布及泊松分布、正态分布、统计抽样、总体和样本的估计、置信区间的构建、假设检验、卡方分布、相关与回归。从开篇至收尾，这些方法由独自登场到联袂表演，从“一支红杏出墙来”到“乱花渐欲迷人眼”，徐徐揭开统计学神秘的面纱。作者充分利用几个活泼有趣的情境，淋漓尽致地发挥深入浅出系列“反复阐述”的特色，举重若轻、有条不紊地完成了这本篇幅长达717页、内容多达15章的作品。

身为译者，我当然有责任全力体现这样一部作品的风采。只是怎样一本译作才是“好”译作？这个议论伴随翻译这个营生的生而生，也许也只能伴随翻译这个营生的亡才会亡。与其孜孜不倦地讨论，不如做点切实的事情，向着“好”的目标接近。

在《深入浅出统计学》的翻译过程中，有两件工作可以斗胆请读者放心：

一是术语的翻译。所有术语都有合理出处，优先采用的参考资料是《GBT 3358.1-2009 统计学词汇及符号 第1部分：一般统计术语与用于概率的术语》。为了方便读者研究，现已将此标准中的所有术语整理成《GB术语表》，见此链接：http://t.cn/aO8vlq。除了少数几个例外，只要能在此标准中查到的术语，均优先使用。所提到的少数几个例外则是为了成全表述上的一致性和逻辑性，例如，“range”一词，未采用GB中出现过的“极差”，而是采用其他资料中出现的另一种说法“全距”，这是为了让与“range”有关的几个术语——全距、四分位距、百分位距等保持和谐一致，便于理解和记忆。

二是标点符号和字体格式的认真使用：第一，统计术语首次出现时加粗，凸显这个术语，引起读者重视；第二，统计用语过于冗长时，为了避免这个用语与句子的其他组成成分混淆不清，也加引号进行提示，例如，第4章第133页上的“停球结果为7”；第三，提高破折号的使用频率，当前后两个句子有较为密切的因果联系时，即使原文未使用破折号，译文也改句号为破折号，以作提示。

遗憾的是，有极少量术语无法做到既坚持采用术语，又保持中文的流畅。这固然是由于我个人经验不足、能力有限，但中英文的习惯差异、技术类书籍对术语正确性和一致性的特定要求也是其中的原因。例如，“expect”译为“期望”，可是，当读到第5章中“这是你每拉一次杆能够期望得到的收益 —— 是负数！”这样的句子时，我们多少会觉得有些怪异 —— 在中文中，我们习惯于将“期望”这个词和美好的前景联系在一起，谁会期望自己的收益是负数？可是，如果将“期望”顺手改为“预期”以满足流畅性的需要，那么从技术的严谨性来看，并不是十分负责的做法。这样的词还有“成功” —— 只要某件事情如期发生，都叫做“成功”，无论好歹。每到这种纠结时分，我往往按照个人意愿，任性地坚持“直译”。衷心希望这样的不足不会影响读者对统计学的兴趣和研究。

本书初稿提交后，在卢鸫翔编辑的热心联络下，以下素未谋面的专业人士参与了初稿试读，给出了宝贵的审定意见，现借译者序向以下热心人士致谢：

陈钢、高涛、江思源、梁涛、林宇、刘思喆、卢涛、孙强宇、魏太云、吴昕、徐旭铭、张伸

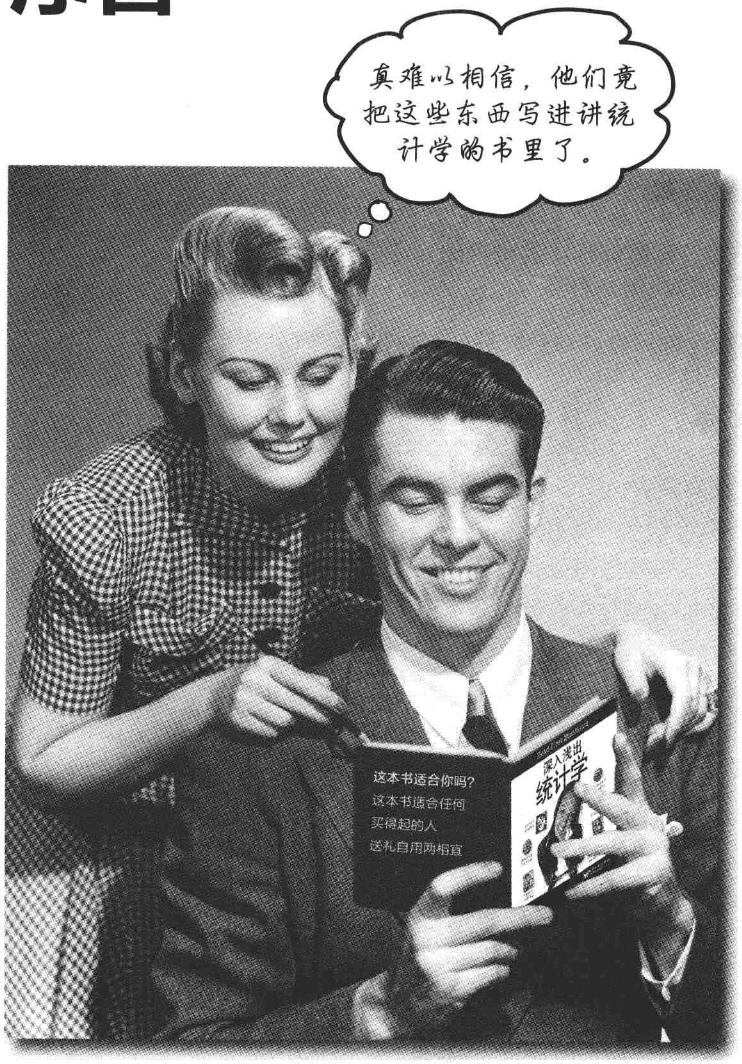
最后：

感谢电子工业出版社、武汉嘉士特文化传播公司对我的信任。
感谢卢鸫翔、徐涛编辑对我的帮助和指教。
感谢家人对我的理解和支持，我爱你们。

李芳

2011年9月

序言



本节回答一个热门问题：“作者为什么非要把这些东西写进一本讲统计学的书里？”

谁适合阅读本书？

请先回答几个问题：

1. 因为一门**课程**、因为自己的**工作**，或仅仅因为你觉得自己**应该懂得标准差的意义或是如何求轮盘赌赢钱概率**，使你需要了解统计学？

2. 你想**学习、理解**和**记忆**如何**使用概率和统计百发百中地得出正确结果**？

3. 与**枯燥**、**无聊**的学术演讲相比，你更喜欢参加充满欢声笑语的晚宴，对吗？

如果以上问题全部回答“对！”— 这本书适合你。

谁该和本书说拜拜？

请先回答几个问题：

1. 你从来没有学过基础算术？（不用很精通，但应该会基本的加、减、乘、除。）

2. 你是一位意气风发的统计师，正在物色参考书？

3. 你惧怕尝试新事物，宁可上山打虎也不愿标新立异，对吗？你认为要是用拟人的手法叙述维恩图，统计学书籍就难免有失严肃，对吗？



只要有一个问题回答“对！”
— 你与本书无缘。

[营销部插话 — 只要有信用卡就可以买书哦。]

我们了解你在想什么

“这怎么能是一本严肃的统计学书籍呢？”

“这些图都是用来干嘛的？”

“我真能这样**学**统计学吗？”

我们了解你的大脑在想什么



你的大脑渴望新事物。大脑总是不停地搜索、探查、**等待**不同寻常的事物，它天生如此，这正是你活力的来源。

那么，大脑怎么对待你所碰到的常规、普通、一般的事情呢？—— 它会竭尽全力阻止这些事情，以免干扰自己**真正的**工作 —— 记录**重要**事项。大脑不会费力保存这些琐事，这些琐事从来不会成功地闯过“明显不重要事项”关卡。

你的大脑如何**知道**哪件事重要？假想有一天你出门旅行，迎面扑来一只吊睛白额大虫，你的头脑和身体会有什么反应？

神经元发动、情绪激动、化学物质激增。于是，你的大脑知道 ——

这事绝对重要！记住！



但，想象你是呆在家里，或是呆在图书馆里，也就是说，是在一个安全、温暖、没有老虎的地方。

你正在复习迎考，要不然就是在努力搞懂一些艰深的技术，你的老板认为花个把星期就能搞定，顶多十天。

唯一的问题是：你的大脑想好好帮你一把，它试图保证不让这种“明显不重要”的内容去破坏珍惜的资源，这些珍稀的资源最好用来保存真正“**重大**”的事情，像老虎啊，像火灾险情啊，像你绝不该在Facebook的网页上贴上那些“聚会”照片啊。

没有什么方便的办法可以告诉大脑“喂，大脑，我对你感激之至，可是不管这本书多无聊，也不管我对这本书多没兴趣，我真的想把这些东西全都记住。”

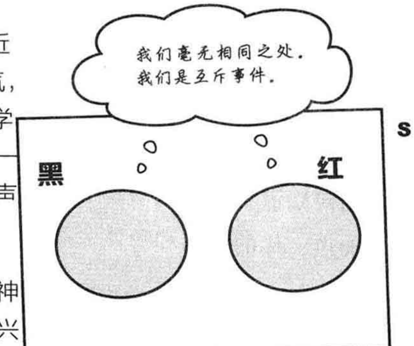
我们认为深入浅出（Head First）系列图书的每一位读者都是一位学习者。

既然要学习，怎样才能学会呢？首先，你得搞懂，然后，切勿遗忘。一字一句硬塞不是办法。根据最新的认知科学、神经生物学及教育心理学研究结果，学习远不仅仅是读书认字。我们知道怎么让你的脑筋动起来。

下面是部分深入浅出（Head First）教学原则：



将知识图形化。图形比单调的文字好记得多，可以提高学习效率（记忆学习和转移学习的学习效率最多能提高89%）；图形还能让知识更容易理解，相比将文字放在页脚和下一页，**将文字放在相关图形当中或图形周围**，学习者成功解决相关问题的可能性将成倍增长。



采用对话式的个性化风格。最近的研究表明，要是回避一本正经的语气，代之以对话般的风格，以第一人称平易近人地给学生上课，学生的课后测验成绩最多可提高40%。多讲几个故事，少来一点高谈阔论，语气宜随和。别太郑重其事。想想看，一局欢声笑语的晚宴，或是一场枯燥的演讲，哪一样更让你感兴趣？

引导读者深入思考：换句话说，除非读者主动调动自己的神经元，否则脑袋里不会发生什么大变化。只有激发读者的兴趣，引起读者的好奇，刺激读者的灵感，读者才能解决问题，得出结论，获得新知识。为此，讲授者要设计各种难题、练习，提出引人深思的提问，还要多设计一些让左右脑半球和多种感官都动起来的活动。



牢牢吸引读者的注意力。大家都有这样的体验——“我是真想学，但看完第一页就晕了”。大脑注意的是不同寻常的、有趣的、奇怪的、引人注意的、出人意料的事情。学习一种新颖艰深的技术不一定非得枯燥乏味不可，果真有吸引注意力的内容，大脑会学得更快。



影响读者的情感。现在我们知道，人的记忆能力在很大程度上取决于要记忆的内容对情感的影响。我们关心什么，就会记住什么；我们对什么事有感觉， 就会记住什么。 这里讲的情感并非天灾人祸给人带来的撕心裂肺的伤痛情感，而是惊讶、好奇、感觉有趣、想追根究底之类的情感，以及在猜对一个字谜、在学会别人感觉难以学会的事情或是在意识到自己懂的东西居然比工程部那位开口闭口“我比你有技术”的张三还多时，油然而生的“我是老大”的感觉。

元认知：对思考的思考

如果真想学东西，而且想学得更快更深入，就要关注自己如何集中注意力。要思考自己的思考方式，研究自己的研究方式。



大多数人在成长过程中都不曾学习元认知和学习理论方面的知识。人们期望我们学知识，但极少有人教我们如何学。

但想象得到，捧着本书的你，的确想学习统计学知识，同时可能不想花费太多时间。要想利用在本书中读到的知识，就得**记住**读过的知识，为此必须**理解**这些知识。为了淋漓尽致地发挥本书或任何书本或学习经验的作用，请管好你的大脑，请管好大脑对待本书的态度。

诀窍在于让大脑把正在学习的新资料当做“正经大事”——对幸福至关重要的大事，像面对老虎一样重要的大事。若非如此，你就会陷入一场持久战：你竭力要记住新知识，大脑却竭力要把这些新知识踢出去。

既然如此，如何让大脑像对待吃人的老虎一样对待统计学知识呢？

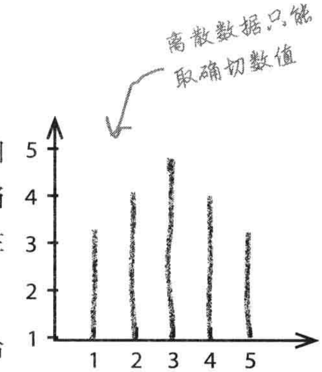
有两种办法，一种缓慢而乏味，一种迅速而有效。慢办法是简单记忆。你显然明白，只要不停地把同样的东西往大脑里灌，即使是最乏味的知识，也能学会、记牢。只要重复灌的次数足够多，大脑就会想：“这些东西给他的**感觉**并不重要，但他不停地看这些相同的东西，**一遍，一遍，再一遍**。因此我猜这些东西肯定很重要。”

快办法是做**一切增进大脑活动**的事，尤其是不同类型的大脑活动。上一页讲了很多这样的活动，事实证明，这些活动全都能促使大脑以有利于己的方式工作。例如，研究表明，将文字放在文字所描述的图片当中（相反的做法是将文字放在页面中的其他位置，如注释位置或正文位置），会促使大脑努力想搞清楚文字和图片之间的关系，进而发动更多神经元。更多神经元发动 = 更有机会让大脑明白某件事值得注意，可能还值得记住。

对话式的写作风格对此很有帮助。人们在与人对话时注意力会更集中，原因是别人期待对方有所表现。令人惊讶的是，大脑不一定会**在意**“对话”是在人和书之间进行！反之，要是写作风格了无新意，乏味枯燥，大脑的感觉就和在挤满消极听众的屋子里听演讲没什么两样：没必要保持清醒。

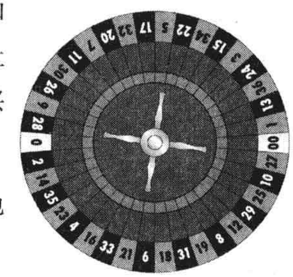
不过，图形和对话式风格只是起步……

我们的做法：



我们使用**丰富的图片**，这是因为，大脑追逐图像，而非文字。在大脑的活动中，一张图片胜过千言万语。当同时使用图片和文字进行说明时，我们将文字填写在图片当中，当文字出现在它所描述的事物当中时，大脑的工作更有效率；相反，若将说明性文字放在注释或其他正文当中，则无此效果。

我们使用**反复论述法**，即以不同的方式、通过不同的媒介对同一主题进行反复描述，给读者营造丰富的感受，目的是让这些主题有更多机会印在大脑的多个区域。

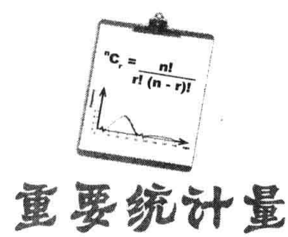


我们以**出人意料的**方式叙述概念和使用图片，因为，大脑追逐新鲜事物；我们在图片和创意中或多或少加入了一些**情感性**的内容，因为，大脑关注情感的生物化学反应。让人有所感触的东西更可能让人记住，即使这点感触不过是一丝幽默、一丝惊讶或一丝兴趣，也是如此。

我们使用个性化的**对话式写作风格**，因为，当大脑认为你是在进行对话而不是在消极地听报告时，就会调整到注意力更集中的状态。即使在读书时，大脑也是这个习惯。



我们安排了80多个**活动**，因为，相比读书，在做事时，大脑经过调整，能学会和记住更多东西。我们安排的练习有难度，但不会让人束手无策，这正是大多数人愿意做的练习。



我们使用**多种教学风格**，因为，有的人可能喜欢一步一步按顺序来，有的人可能喜欢先看懂大图，还有一些人可能只想看看例子。我们将以多种方式反复讲述相同的主题，不管读者的个人爱好如何，都将因此受益匪浅。

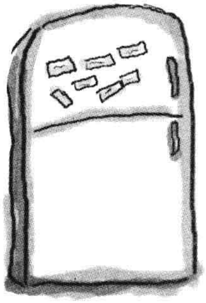
我们安排了让**左右脑半球**分别负责的内容，因为，大脑开动部位越多，就学得越多，记得越多，注意力更持久。由于一侧大脑工作往往意味着另一侧大脑得到休息，左右半脑的分工合作使得长时间学习的学习效率得到提高。

我们还安排了一些**场景**和练习，在场景中展现**不同的观点**，因为，当大脑被迫进行评估和判断时，会调整到深入学习状态。



我们在练习中安排了一些**难点**，即提出一些无法简单回答的**问题**。因为，你的大脑在不得不处理某件事情时，会调整到学习和记忆状态。开动脑筋吧，“光看别人做运动无法让自己体态健美”。别担心，我们尽力保证，你努力学习的都是该学的，你不会为了对付一个费解的例子或为了分析一段用词过于晦涩或行文过于简化的段落而**多用一个脑细胞**。

我们以**人物**为例，把人物安排在场景、实例、图片等等内容中。至于原因嘛，因为你是人群中的一员啊，你的大脑对人比对事更关注。



你的任务：征服大脑

我们的工作到此为止，剩下的就看你的了。从下面这些提示出发，顺从大脑的判断，看看哪些对你有用，哪些对你没用，尝试一下新事物吧。

① **慢慢读。理解的内容越多，要记忆的内容越少。**

忌**死读**。停一停，想一想，碰到书中的提问时，别直接翻看答案；想象真的有人在问你这个问题。强迫自己的大脑想得越深，学会、记住的概率就越大。

② **自己做练习，自己记笔记。**

我们安排了练习和笔记，但是，要是我们替你完成，就像让别人替你锻炼身体一样。只动眼不动手也不可取，要**动动笔**。大量证据证明，学习时的身体动作能提高学习效率。

③ **阅读“世上没有傻问题”部分**

世上没有傻问题。这些问题并非可看可不看，**这是核心内容的组成部分**！请勿忽略。

④ **请将下面这段话作为最后一段床头文字，或起码作为最后一段高深的床头文字。**

有一部分学习过程（尤其是短暂记忆转变为长期记忆的过程）发生在放下书本**之后**，大脑需要有自己的时间进行更多处理。如果在这段处理时间内学新东西，将会丢失一些刚学会的东西。

⑤ **大量喝水。**

充沛的体液会让大脑处于最佳工作状态，脱水（早在感到口渴前就会发生）则会让认知功能下降。

⑥ **开口大声讨论。**

说话会刺激大脑的其他部分。如果你正在努力理解一些知识，或正在努力增加以后记住这些知识的概率，请大声说出这些知识。还有一种更好的做法，试着向别人大声解释这些知识。你会学得更快，可能还会发现一些阅读时不曾发现的名堂。

⑦ **聆听大脑的声音。**

留意你的大脑是否超负荷工作。若你发现自己开始心不在焉，或是刚刚读过的东西转眼忘记，就该休息。一旦过了某个学习点，哪怕拼命塞，也无法提高学习效率，反而有可能影响学习过程。

⑧ **找到感觉。**

大脑需要知道事情是否重要。让自己融入各种场景，为照片设想旁注，就连抱怨一个并不好笑的玩笑，也比什么感觉都没有强。

⑨ **勤加练习！**

完全掌握统计学的唯一办法就是勤加练习，本书通篇都在引导你这样做。统计学的运用是一门技术，精于此道的唯一办法就是大量练习。本书将为你提供大量练习机会：每一章中都有一个等待你解决的问题，千万别跳过这些问题不看 —— 大量学习都发生在解决问题的过程中。我们为每一个问题提供了答案，要是卡了壳（有些细微之处很容易给人带来麻烦），别不敢看！不过，请尽量先解决问题，然后再看答案，务必让你的办法行之有效，然后才继续看书中的下一部分内容。

本书自述

本书是经验之谈，并非系统教材，我们故意抽掉了会妨碍讲述书中相关知识的东西。本书对你已经见识过和学习过的知识作了一些假设，因此第一次通读本书的时候，需要从头读起。

我们从数据展示和汇总的基本方法讲起，讲至概率分布，再讲至假设检验等其他高级应用技术。

尽管越往后的章节越重要，但起始部分却是后文的基石，图表绘制、均值计算，以及变异性的量度均在此列。因此，我们一开始就抛给你一些基本统计问题，这些问题其实是靠你自己解决的。如此一来，你就能立刻把统计学用起来，你将开始摩拳擦掌。在此之后，我们继续向你讲解概率和概率分布的运用，这时你将打下坚实的统计学基础，可以专心于学习概念。接着，我们讲解如何挥洒自如地运用所学知识，假设检验就是一例。我们瞄准时机为你提供合适的知识，这样的知识最有价值。

我们所讲范围与AP与A Level的常规范围相同。

尽管我们的关注重点是全面学习而非考试准备，但我们很好地纳入了AP和A Level课程的内容，因此，随着对本书的学习，你将深入领会各种等级考试的应知应会内容，通过这种方法学习统计学比死记硬背公式效果更好，因为你对于自己的需要充满自信，你知道什么时候需要使用这些知识，知道如何使用这些知识。

我们通过在线资源为你提供帮助。

通过与读者交流，我们得知你有时候需要一些特别的帮助，因此我们提供了一些在线资源，动动指尖就能访问。我们为你提供互动在线论坛、在线论文及其他资源。请从这里开始访问：

http://www.headfirstlabs.com/books/hfstats/

活动并非可做可不做。

练习和活动不是点缀，而是本书的核心组成部分。这些练习和活动有的是为了帮助记忆，有的是为了帮助加深理解，还有的是为了帮助应用所学知识，**切勿忽略**。唯一的例外是填字游戏，这不是非做不可，不过，填字游戏会让大脑得到在不一样的语境下思考所学到的单词和术语的大好机会。

反复论述是刻意而重要的安排。

深入浅出系列书籍有一个明显特色：我们希望你“真正”掌握学到的知识，我们希望你在看完本书的同时就记住学到的知识。大多数参考书都不把记忆和回忆当做一个目标，但本书的目标是“学会”，所以，常常会看到同一概念多次出现。

“动动脑”练习没有答案。

有一些“动动脑”练习没有标准答案，另有一些练习是为了让你判断自己的答案是否正确，以及在什么情况下会正确。部分“动动脑”练习给出了提示，为你指明正确方向。

技术顾问组

Ariana Anderson



Cary Collett



Dru Kleinfeld



Danielle Levitt



Andy Parker



Michael Prerau



无照片（同样很出色）

杰弗里·马德勒纳（Jeffrey Maddelena）与马修·维德朋克（Matt Vadeboncoeur）

同时感谢凯斯·法尔格伦（Keith Fahlgren）、布鲁斯·弗雷（Bruce Frey）和利安娜·洛克哈特（Leanne Lockhart）的技术反馈。

技术顾问：

阿瑞娜·安德森（Ariana Anderson），加利福尼亚大学洛杉矶分校统计系教师助理及博士生，她的研究兴趣包括数据挖掘和模式识别。

德鲁·克莱因菲尔德（Dru Kleinfeld），康奈尔大学2007届毕业生，经济学学士。德鲁目前住在纽约，供职于摩根斯坦利人力资源部。

丹妮尔·莱维特（Danielle Levitt），目前在以色列特拉维夫学医，为一年级学生。闲暇时候，她欣赏音乐、游海泳、花时间陪伴朋友和家人。

安迪·帕克（Andy Parker），以物理学研究员的身份开始了他的职业生涯，他认为自己对统计学略知一二。读完这本书后，他感到深受启发。Andy现在大部分时间都在担心自己是不是还忘掉了什么别的重要事情。

迈克尔 J. 普瑞诺（Michael J. Prerau），计算神经科学研究员，他创造了一种统计方法，用于分析大脑中的神经元如何对信息进行编码。他是波士顿大学神经科学编程专业的博士生，同时也是麻省综合医院麻醉学与重症看护科神经科学统计学研究实验室的助理研究员。

马修·维德朋克（Matthew Vadeboncoeur），新罕布什尔大学复杂系统研究中心生态系统生态学研究生。

致谢

我的编辑：

衷心感谢我的编辑山德斯·克林菲尔德（Sanders Kleinfeld）。与山德斯合作是种享受，通过电话、电子邮件、聊天，主顾变成了朋友。他的投入和刻苦令人惊叹，他为我提供的建议和支持堪称一流。谢谢，山德斯！你棒极了，没有你我无法写出这本书。



Sanders Kleinfeld

O'Reilly团队：

超级感谢布雷特·麦克劳林（Brett McLaughlin），是他让我飞到波士顿参加Head First新手营，给了我一生难得的机会。布雷特对Head First的直觉非常到位，我十分感谢他在本书写作过程中给我的指导。谢谢，布雷特。



Lou Barr

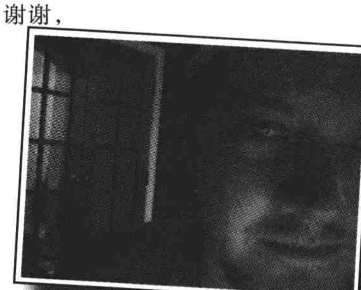
这本书要是没有露·巴尔（Lou Barr），难以想象会是什么样子。露是一位令人惊叹的图形设计师，她为本书妙笔添花，不仅如此，她让整个学习体验更具内涵。她举重若轻，我对她感激不尽。

Head First团队的其他成员也当得起一声郑重的谢谢。凯瑟琳·诺兰（Catherine Nolan）扶助我度过了这本书的前期阶段，让我在Head First天地里无拘无束，布列塔尼·史密斯（Brittany Smith）保持生产稳定运行，劳琳·帕翠基（Laurie Petrycki）给了我充分的信任，把这本书交给我来写（她在年初还把办公室借给我用 —— 谢谢，劳琳！），凯斯·麦克纳马拉（Keith McNamara）出色地组织了一支技术审核团队，凯琴·麦卡洛（Caitrin McCullough）打理网站的一切事务。多谢各位！



Catherine Nolan

我要特别提一提凯西·赛拉（Kathy Sierra）和伯特·贝茨（Bert Bates），他们是这一系列精彩图书的原始策划者。本书能够成为这个系列中的一环不胜荣幸。



Brett McLaughlin

我的家人和朋友：

真希望有足够的篇幅把所有帮过我的人都写上一笔，特别要感谢大卫（David）、妈妈、爸爸、史蒂夫·哈维（Steve Harvey）、吉尔·切斯特（Gill Chester）、保罗·伯吉斯（Paul Burgess）、安迪·塔特勒（Andy Tatler）以及彼得·沃克（Peter Walker），是你们鼓励我前进，对你们的支持和鼓励我万分感激，无以言表，谢谢你们。

Safari®在线图书馆

若你在自己喜爱的技术书籍封面上看到Safari图标，说明该书已在O'Reilly网络图书馆（O'Reilly Network Safari Bookshelf）上架销售。

Safari为读者提供比电子书更胜一筹的解决方案，它是一个虚拟图书馆，你可以轻松地搜索成千上万本顶尖技术书籍，剪贴代码样本，下载图书章节，在需要获得最精确、最先进的信息时迅速找到答案。请免费试用：

`http://safari.oreilly.com`

总目录

	序言	xxvii
1	信息图形化：第一印象	1
2	集中趋势的量度：中庸之道	45
3	分散性与变异性的量度：强大的“距”	83
4	概率计算：把握机会	127
5	离散概率分布的运用：善用期望	197
6	排列与组合：排序、排位、排	241
7	几何分布、二项分布及泊松分布：坚持离散	269
8	正态分布的运用：保持正态	325
9	再谈正态分布的运用：超越正态	361
10	统计抽样的运用：抽取样本	415
11	总体和样本的估计：进行预测	441
12	置信区间的构建：自信地猜测	487
13	假设检验的运用：研究证据	521
14	χ^2分布：继续探讨……	567
15	相关与回归：我的线条如何？	605
附录i	尾声：正文未及的十大拓展	643
附录ii	统计表：快来查表	657

细分目录及各章引子

序言

大脑对待统计学的态度。一边是你努力想学会一些知识，一边是你的大脑忙着开小差。你的大脑在想：“最好把位置留给更重要的事，像该离哪些野生动物远点啊，像光着身子滑雪是不是个坏点子啊。”既然如此，你该如何引诱你的大脑意识到，懂得统计学是你安身立命的根本？

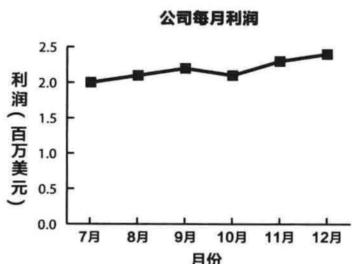
谁适合阅读本书？	xxx
我们了解你在想什么	xxxi
元认知	xxxiii
征服大脑	xxxv
本书自述	xxxvi
技术顾问组	xxxviii
致谢	xxxix

1 信息图形化

第一印象

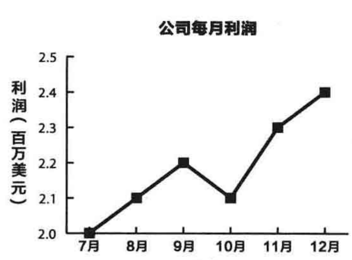
在为手头数据无法给出事情真相而发愁吗？统计能**化繁为简**，帮助你让一堆堆令人困惑的数据发挥作用。当你发现数据的真相后，接下来就需要借助**可视化**的方法表现出来，使之**公之于众**。为了找到最合适的图表完成这个过程，请你整理衣衫，带上最好的计算尺，和我们一起赶往“统计邦”吧。

统计量无处不在	2
为何学习统计学？	3
从两张图说起	4
呆板的饼图	8
条形图更具精确性	10
垂直条形图	10
水平条形图	11
标度的影响力	12
使用频数标度	13
处理多批数据	14
类别与数字	18
处理分组数据	19
绘制直方图起步：求出长方形宽度	20
第1步：求长方形宽度	26
第2步：求长方形高度	27
第3步：画出直方图	28
认识累积频数	34
绘制累积频数图	35
选择正确的图形	39









2 集中趋势的量度

中庸之道

有时候，把握问题核心才是当务之急。从一大堆数字中看出模式和趋势可能颇为不易，而求出**平均数**往往是把握全局的第一步。有了平均数就能迅速找出数据中最具代表性的数值，得出重要结论。在本章中，我们将介绍几种方法，帮助你计算最重要的统计量——均值、中位数、众数。你将开始学习如何有效地**汇总数据**，尽可能得出简练、有用的结果。

健身俱乐部
统计邦顶级养生馆

欢迎来到健身俱乐部	46
均值：平均数的一般量度	47
均值数学	48
处理未知条件	49
再说均值	50
再说健身俱乐部	53
人人都在练功夫	54
我们的数据中存在异常值	57
真凶是异常值	58
饮水机边的对话	60
寻找中位数	61
求中位数三步法：	62
生意日益兴隆	65
小鸭呱呱游泳班	66
均值和中位数出了什么问题？	69
我们该怎么处理这样的数据呢？	69
均值访谈	71
认识众数	73
求众数三步法	74

20岁



20岁



21岁





20岁



19岁



3 分散性与变异性的量度

强大的“距”

世事可靠不可靠，我们该问谁？平均数在寻找数据集典型值方面十分了得，**但平均数并不能说明一切。**平均数能让你知道数据中心所在，但若要给数据下结论，仅有均值、中位数和众数往往无法提供充足信息。在本章中，我们将开始分析**各种距和差**，让你的数据分析技术进入新境界。



招聘：队员一名 84
我们需要比较球员得分 85
使用全距区分数据集 86
异常值带来的问题 89
我们需要摆脱异常值 91
四分位数出手相救 92
四分位距剔除异常值 93
剖析四分位数 94
我们并不局限于使用四分位数 98
什么是百分位数？ 99
用箱线图绘制各种“距” 100
变异性比分散性更具体 104
计算平均距离 105
我们可以用方差计算变异性…… 106
但标准差才是更直观的量度方法 107
标准差访谈 108
方差速算法 113
碰上需要比较基准的情况该怎么办？ 118
使用标准分比较不同数据集中的数值 119
标准分释义 120
统计邦全明星篮球队赢了联赛！ 125



4 概率计算

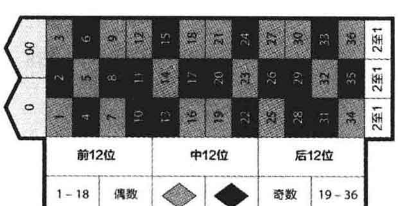
把握机会

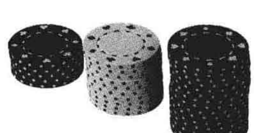
人生无常瞬息之间的变化有时难以一一料定。但有些事情会比其他事情更有可能发生，这就为**概率理论**提供了大显身手的舞台。通过概率能评估出现各种结果的可能性，让你**预测未来**。知悉可能出现的结果则可帮助你作出**有根据的决策**。本章将让你了解更多概率知识，学会如何掌控未来！



肥蛋大满贯	128
转起来吧，轮盘！	129
几率有多大？	132
求解轮盘概率	135
维恩图：概率的图形表示	136
你还可以将几个概率相加	142
互斥事件与相交事件	147
交集带来的问题	148
更多表示法	149
又一次倒霉的转动……	155
设定条件	156
求解条件概率	157
利用概率树还能计算条件概率	159
概率树使用诀窍	161
第1步：求P(黑∩偶)	167
第2步：求P(偶)	169
第3步：求P(黑|偶)	170
利用全概率公式求解P(B)	172
认识贝叶斯定理	173
如果几个事件互有影响，则为相关事件	181
如果几个事件互不影响，则为独立事件	182
再谈独立事件概率计算	183







5 离散概率分布的运用

善用期望

意外从天而降，未来如何演变？前文讲到如何通过概率得知发生某些事件的可能性的大小。可惜概率并非万能，它无法指出所发生的这些事情的**整体影响**，也无法指出这种整体影响对你的具体影响。不错，你有时会在轮盘赌中大赚特赚，但你赚到的钱真的填得平那些赔掉的钱吗？在本章中，我们将讲述如何利用概率**预测长期结果**，以及如何量度这些**预测结果的确定性**。

重回肥蛋赌场 198

我们可以写出老虎机概率分布 201

期望指示预测结果…… 204

方差指示结果的分散性 205

方差和概率分布 206

让我们算算老虎机的方差 207

肥蛋改了价码 212

E(X)与E(Y)之间存在线性关系 217

老虎机变换 218

线性变换的通用公式 219

每一次拉杆为一个独立观测值 222

观测值速算法 223

新老虎机在等你 229

E(X) + E(Y)= E(X + Y) 230

E(X) − E(Y)= E(X − Y) 231

线性变换也可以做加减运算 232

发了！ 238



6 排列与组合

排序、排位、排

顺序有时很重要 —— 清点某些事物的**所有可能排序方法**耗时颇巨，可这却是计算某些概率**必不可少**的过程 —— 麻烦就在这里。在本章中，我们将介绍推导出这类信息的**简便方法**，为你免除清点一切可能结果的烦恼。来吧，让我们看看如何**计算概率**。



统计邦德比杯马赛 242

三马赛正在进行 243

马儿们有几种穿越终点线的方式？ 245

计算排位数目 246

圆形排位 247

花样赛开始了 251

按个体排名与按种类排名不是一回事 252

我们需要按种类排列动物 253

推导出用于重复排列的公式 254

二十马赛正在进行 257

前三甲归属方式有几种？ 258

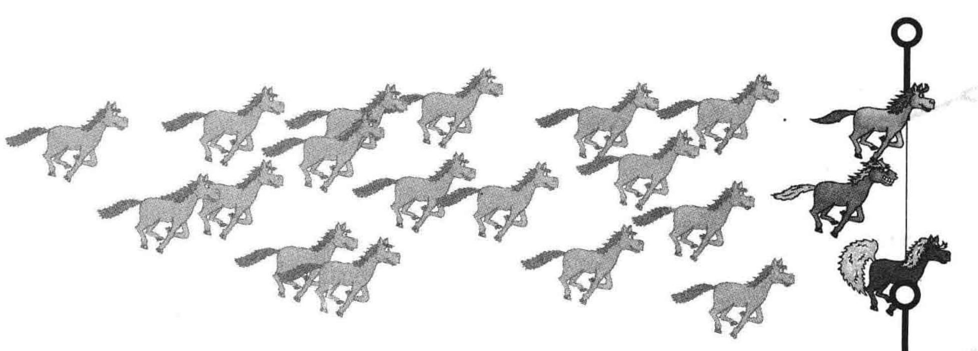
何为排列 259

假如马匹排名无关紧要 260

何为组合 261

组合访谈 262

比赛结束 268



7 几何分布、二项分布及泊松分布

坚持离散

计算概率分布颇为费时。前面讲到如何计算和利用概率分布，不过，如果**方法更简单一些**，**计算速度更快一些**，效果岂不更好？在本章中，我们将介绍一些**特殊的概率分布**，这些概率分布有着十分固定的模式。只要懂得这些模式并善加利用，就能以前所未有的速度**计算概率、期望、方差**。接着读吧，让我们一起来认识几何分布、二项分布及泊松分布。

我们需要求出查德的概率分布	273
这种概率分布有一种固定模式	274
概率分布可以用代数式表示	277
几何分布对不等式同样有用	279
几何分布的期望模式	280
期望是 1/P	281
求当前分布的方差	283
几何分布简明指南	284
转椅赢赢赢！	287
你已经掌握了几何分布	287
玩下去，还是转身走？	291
推广到求3个问题的概率	293
进一步推导概率算式	296
期望和方差如何计算？	298
二项分布的期望与方差	301
二项分布简明指南	302
泊松分布的期望和方差	308
概率分布是怎样的？	312
组合泊松变量	313
伪装下的泊松分布	316
泊松分布简明指南	319







8 正态分布的运用

保持正态

离散概率分布并非无所不能。到目前为止，我们接触到的都是可以指定确切数值的概率分布。然而并非所有数据集合都是如此，还有几类数据**并不符合**我们之前遇到的概率分布。我们将在这一章里讲解所谓的**连续型概率分布**，并介绍最重要的概率分布类型之一 —— **正态分布**。



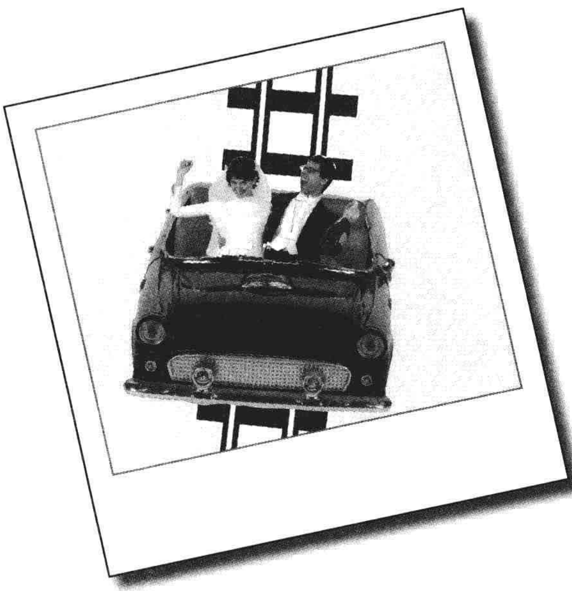
离散数据可取确切值…… 326
但并非所有数值型数据都是离散的 327
推迟几分钟？ 328
我们需要求连续数据的概率分布 329
概率密度函数可用于描述连续数据 330
概率 = 面积 331
欲算概率，先求f(x)…… 332
再求面积，可得概率 333
概率算好了 337
寻找灵魂伴侣 338
男伴模型 339
正态分布是连续数据的“理想”模型 340
如何求正态概率？ 341
正态概率计算三步法 342
第1步：确定分布 343
第2步：标准化为N(0, 1) 344
欲完成标准化，先移动均值…… 345
然后收窄 345
现在，为要计算其概率的特定数值求出Z 346
第3步：用方便易用的概率表查找概率 349



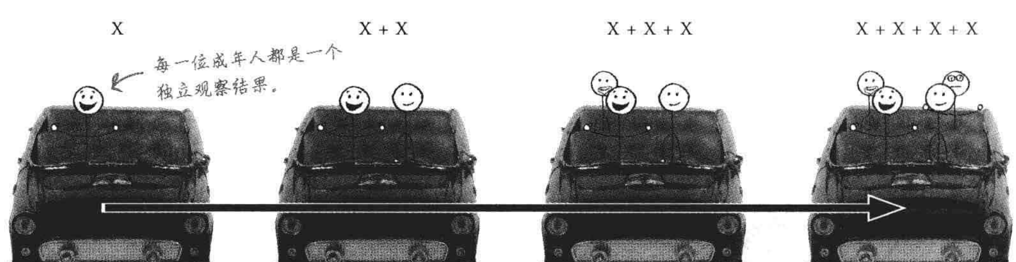
9 再谈正态分布的运用

超越正态

但愿所有的概率分布都是正态分布。有了正态分布，日子**好过多了**——既能一口气查出整个范围的概率，又能留下点时间玩游戏，谁还会花时间一个一个地计算概率呢？在本章中，你将学习如何闪电般**解决更复杂的问题**，还将懂得如何将正态分布的便利运用到**其他概率分布**上。



双双登上爱情过山车	363
正态新娘 + 正态新郎	364
终究还是体重问题	365
综合体重符合哪种分布?	367
求解概率	370
更多人想坐爱情过山车	375
线性变换描述了数据的基本变化……	376
而独立观察结果描述的是你有多少数值	377
独立观察结果的期望和方差	378
接着玩，还是转身走?	383
正态分布出手相救	386
何时用正态分布近似代替二项分布	389
再谈正态近似	394
二项分布是离散分布，正态分布则是连续分布	395
在计算近似值之前先进行连续性修正	396
组合访谈	404
大家坐上爱情过山车	405
何时用正态分布近似代替泊松分布	407
婚礼成功!	413

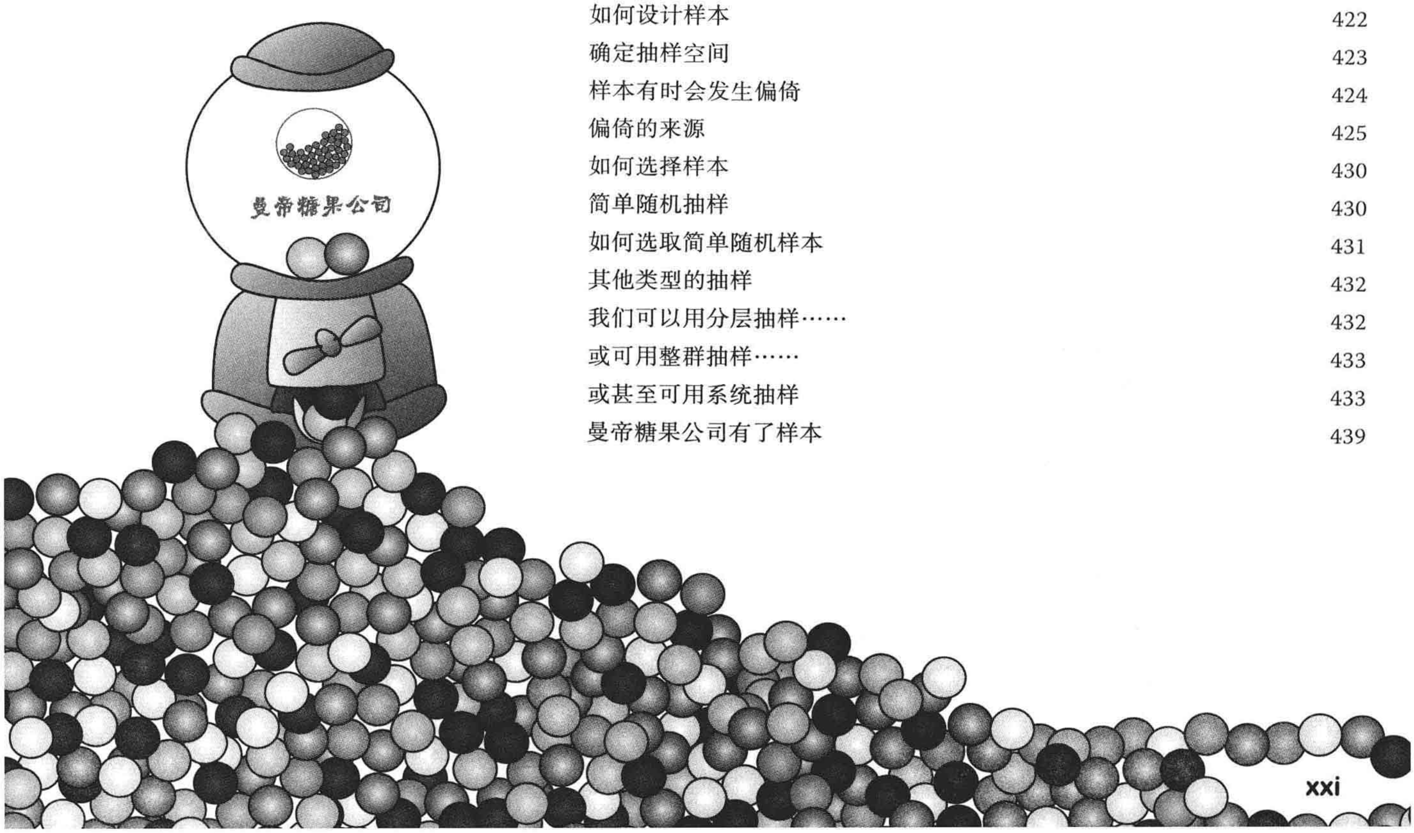


10 统计抽样的运用

抽取样本

统计需要处理数据，数据从何而来？ 有时候数据很容易收集 —— 例如参加一家健身俱乐部的人员的年龄，或一家游戏公司的销售数据；但有时候不太容易，这时候该怎么办？ —— 当事件数量十分庞大时，很难决定该从何处着手收集数据。在本章中，我们将看看如何在实际工作中**成功收集数据** —— 有效地、正确地、省时省钱地收集数据。欢迎来到抽样天地。

曼帝糖果公司口味检验 416
糖球吃光了 417
对糖球样本而非糖球总体进行检验 418
抽样方法 419
当抽样有误时 420
如何设计样本 422
确定抽样空间 423
样本有时会发生偏倚 424
偏倚的来源 425
如何选择样本 430
简单随机抽样 430
如何选取简单随机样本 431
其他类型的抽样 432
我们可以用分层抽样…… 432
或可用整群抽样…… 433
或甚至可用系统抽样 433
曼帝糖果公司有了样本 439



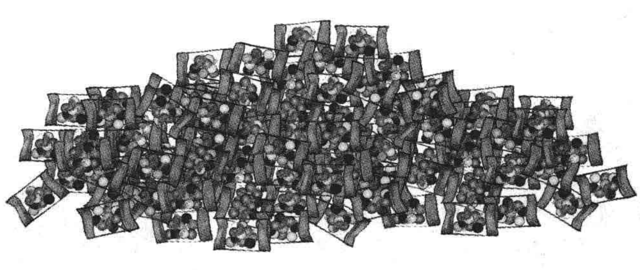
11 总体和样本的估计

进行预测

得样本而知总体，不亦乐乎？若想成为**样本专家**，首先要懂得如何最有效地利用到手的样本——利用样本**准确地预测**总体，并以一定方式说明预测结果的**可靠程度**。在本章中，我们将讲解如何通过样本**了解总体**，以及如何通过总体了解样本。

糖球口味到底能持续多久？ 442
让我们首先估计总体均值 443
点估计量可以近似总体参数 444
让我们估计总体方差 448
我们需要一个有别于样本方差的点估计量 449
哪个公式用在哪里？ 451
这是一个比例问题 454
这和抽样有什么关系？ 459
比例的抽样分布 460
P_s的期望是多少？ 462
P_s的方差是多少？ 463
求解P_s的分布 464
P_s符合正态分布 465
我们需要求样本均值的概率 471
均值的抽样分布 472
求$\overline{X}$的期望 474
$\overline{X}$的方差是多少？ 476
$\overline{X}$如何分布？ 480
当n很大时，$\overline{X}$仍然可以用正态分布近似 481
使用中心极限定理 482



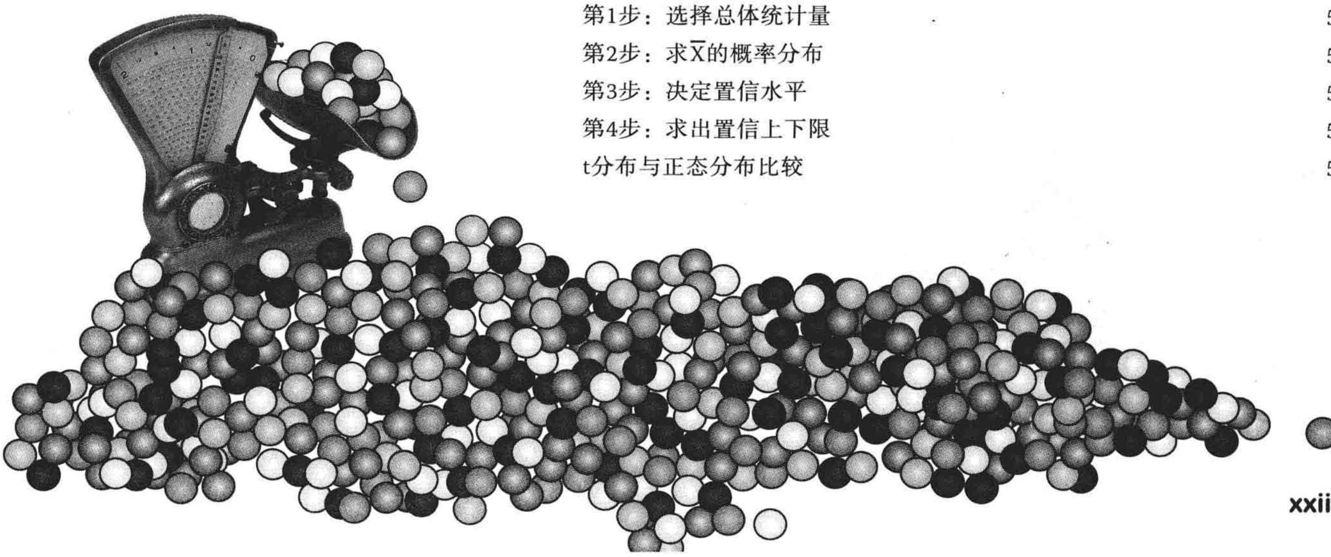


12 置信区间的构建

自信地猜测

有时候样本无法给出足够正确的结果。前面讲到如何用点估计量估计总体均值、方差或一定比例的**精确值**。问题在于，你怎么能肯定自己的估计完全正确？毕竟，你仅仅依靠一个样本对总体作出假设，如果这个样本出问题怎么办？本章将介绍**另一种估计总体统计量的方法**——一种**考虑了不确定性**的方法。拿出你的概率表，我们将向你讲解**置信区间**的来龙去脉。

曼帝糖果出事了 488
精度引起的问题 489
认识置信区间 490
求解置信区间四步骤 491
第1步：选择总体统计量 492
第2步：求出所选统计量的抽样分布 492
第3步：决定置信水平 494
第4步：求出置信上下限 496
先求Z 497
用μ改写不等式 498
最后求$\overline{X}$的数值 501
你求出了置信区间 502
步骤总结 503
置信区间简便算法 504
第1步：选择总体统计量 508
第2步：求$\overline{X}$的概率分布 509
第3步：决定置信水平 512
第4步：求出置信上下限 513
t分布与正态分布比较 515



13 假设检验的运用

研究证据

他人的言论未必句句真实可信。问题是如何判断他人的言论何时真，何时假？假设检验为你提供了一种方法——利用**样本检验**各种统计断言是否可能属实。通过假设检验可以**权衡证据**，检验极限结果——是**纯属巧合**，还是存在其他内在根据？让我们一起阅读本章，看看如何利用假设检验证实或打消你内心深处的疑虑。

统计邦新上市的神奇药品 522
纵观全局 526
假设检验六步骤 527
第1步：确定假设 528
第2步：选择检验统计量 531
第3步：确定拒绝域 532
第4步：求出P值 535
第5步：样本结果位于拒绝域中吗？ 537
第6步：作出决策 537
如果样本增大会怎么样？ 540
让我们再进行一次假设检验 543
第1步：确定假设 543
第2步：选择检验统计量 544
在我们的检验统计中用正态分布近似二项分布 547
第3步：求出拒绝域 548
让我们从第一类错误讲起 556
再谈第二类错误 557
发现鼾克检验的错误 558
我们需要求数值范围 559
求P(第二类错误) 560
认识功效 561



14 χ^2分布

继续探讨……

有时候事实与期望并不相符。当以一种特定的概率分布为某种情况建模时，对于事物的长期可能结果，你有十分清晰的想法。可如果**期望与事实**存在差别呢？你该如何判断？——这些偏差是正常波动，还是说明概率模型存在问题？本章将讲解如何利用 χ^2分布**分析结果**，排除**可疑结果**。

肥蛋赌场可能有麻烦	568
让我们从老虎机开始	569
用χ^2检验评估差异	571
检验统计量代表什么？	572
χ^2分布的两个主要用途	573
ν 表示自由度	574
显著性是多少？	575
χ^2假设检验	576
你解开了老虎机之谜	579
肥蛋遇到了新问题	585
χ^2分布可以检验独立性	586
可用概率求出期望频数	587
频数是多少？	588
我们还需要计算自由度	591
自由度计算方法归纳	596
得出算式……	597
你救了肥蛋赌场	599



15 相关与回归

我的线条如何？

你是否曾经为某两件事的相互关系困惑不已？前面讲过的统计量只描述一个变量——如个人身高、篮球队员得分或是糖球口味持续时间，但是，另外还有一些统计量可以说明**变量之间的关系**。了解事物的相互关系可以丰富你的信息，让你了解真相，使你立于不败之地。来吧，让我们为你介绍**发现事物关系的秘诀**：相关与回归。



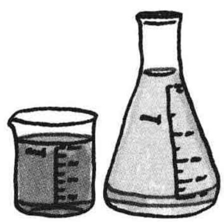
让我们分析天晴时数和听众人数 607
数据类型探讨 608
二变量数据可视化 609
散点图为你指出模式 612
相关关系与因果关系 614
用最佳拟合线预测数值 618
最佳猜测仍是猜测 619
我们需要将误差最小化 620
认识误差平方和 621
求最佳拟合线公式 622
求最佳拟合线斜率 623
求最佳拟合线的斜率，第二部分 624
b求出来了，a呢？ 625
你已经找出了关系 629
让我们查看一些相关关系 630
用相关系数量度直线与数据的拟合度 631
相关系数r有专用计算公式 632
求音乐会数据的r 633
求音乐会数据的r（续） 634



附录i：尾声

正文未及的十大拓展

正文既已，余兴未尽。我们觉得还有一些内容是你需要知道的，对这些内容只字不提恐有不妥，不过，其实也只需要简单地提一提 —— 我们诚挚地希望为你呈上一本厚薄适度的书，免得你为了捧起这本书学习还得先去健身中心练练臂力。因此，请先通读一遍这里的**知识点**，再合上本书。



#1. 数据的其他表现形式 644

#2. 分布剖析 645

#3. 实验 646

#4. 最小二乘回归法的其他公式 648

#5. 决定系数 649

#6. 非线性关系 650

#7. 回归线斜率的置信区间 651

#8. 抽样分布 — 两个均值之间的差异 652

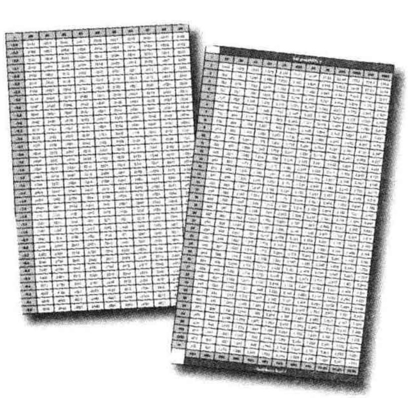
#9. 抽样分布 — 两个比例之间的差异 653

#10. 连续概率分布的E(X)和Var(X) 654

附录ii：统计表

快来查表

缺少值得信赖的概率表该怎么办？仅仅了解概率分布是不够的，有时还需要在标准概率表中**查找概率**。这份附录给出了**正态分布、t分布**和**χ^2分布**的概率表，可在其中尽情查找各种概率。



标准正态分布表 658

t分布临界值 660

χ^2临界值 661

1 信息图形化

第一印象

我想打扮得干干净净、漂漂亮亮，给人留下好印象。



在为手头数据无法给出事情真相而发愁吗？

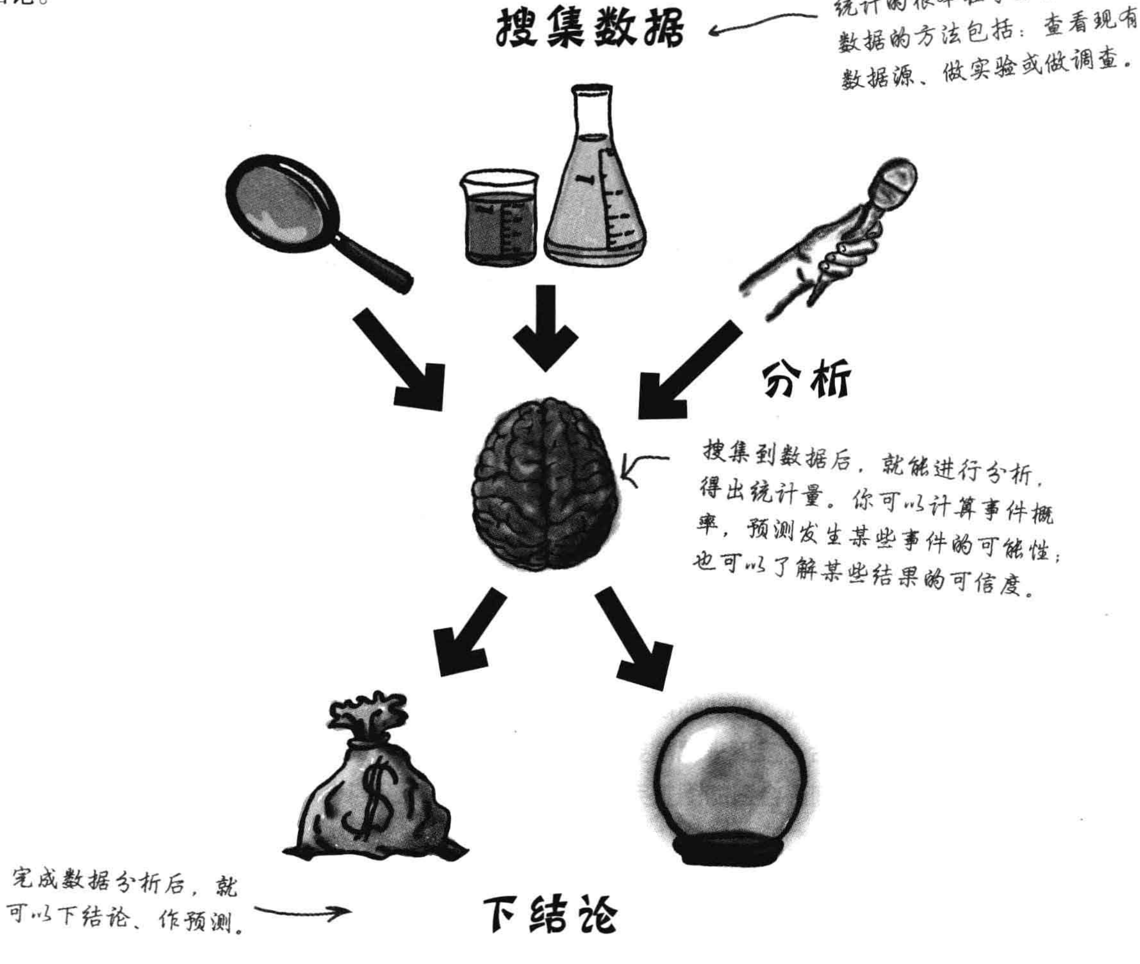
统计能**化繁为简**，帮助你让一堆堆令人困惑的数据发挥作用。当你发现数据的真相后，接下来就需要借助**可视化**的方法表现出来，使之**公之于众**。为了找到最合适的图表完成这个过程，请你整理衣衫，带上最好的计算尺，和我们一起赶往“统计邦”吧。

统计无处不在

网页浏览、运动竞技、游戏排名，但凡人们目光所及，处处皆有统计量。然而，究竟**何为**“统计”？

统计是这样一些数字：它们通过某种有意义的方式对原始事实和数字进行提炼，使得仅仅通过观察原始数据无法立即水落石出的一些理念得以昭示。这里的数据指的是我们能够据其做出结论的事实或数字。例如，若你只想知道自己心爱的球队在联赛中排名如何，大可不必辛辛苦苦地过目诸多赛事的得分记录，只需一个统计量，就能立即得到所需要的信息。

对统计的**研究**包括：统计数据的来源、计算方法及有效使用方法并得出结论。



为何学习统计学？

借助统计方法了解事实真相会令你能力过人，身手不凡。只要**得到**可靠的统计量，就能作出客观的决策，如有神助地进行精确地预测，以及以最有效的方式传达自己想传达的信息。

统计可以成为提炼数据本质的一件法宝，然而也有需小心提防之处。

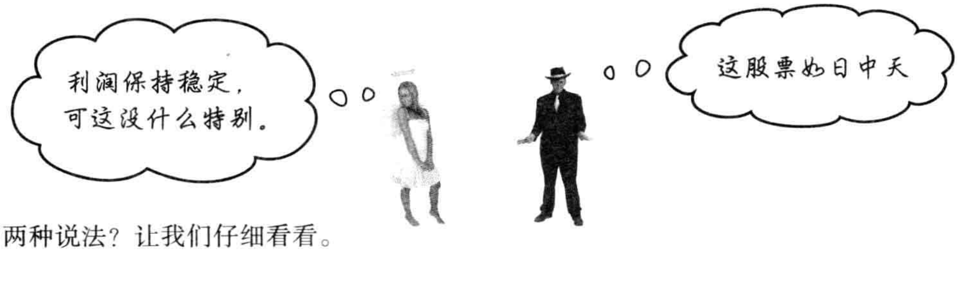


统计以事实为基础，尽管如此，有时却具有误导性。利用统计，既可以昭告事实，也可以瞒天过海。问题是，如何才能判别自己所获悉的是事实，亦或是谎言？

好好掌握统计学将会使你处于有利地位，你将拥有更好的手段去判断统计量是否出错或产生了误导。换句话说，学习统计学是避免遭人愚弄的良策。

请看实例：某公司去年下半年盈利情况。

月份	7月	8月	9月	10月	11月	12月
利润（百万）	2.0	2.1	2.2	2.1	2.3	2.4

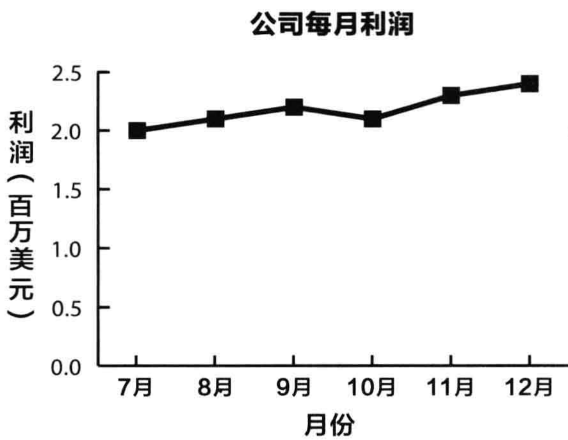


对同一批数据为何会有两种说法？让我们仔细看看。

从两张图说起

我们该怎么探讨针对同一批数据的这两种不同解释呢？—— 我们需要用某种方式直观地表现这些数据。说到信息的直观表现形式，最好的方法莫过于图表。图表是概括原始信息的便捷方式，能帮助你一眼得出初步印象。不过要小心，即使最简单的图表也能神不知鬼不觉地迷惑你、误导你。

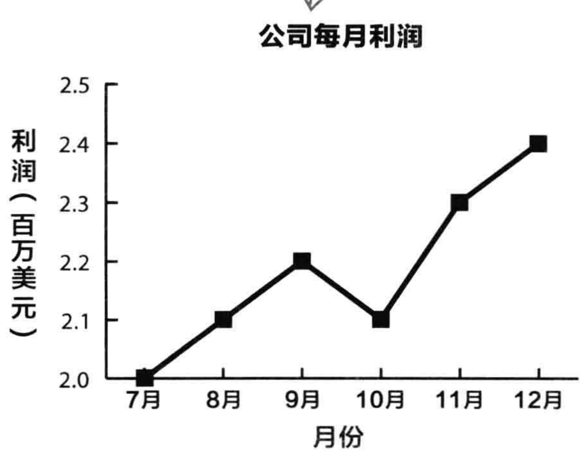
下面这两张图体现了某公司6个月的赢利情况。它们都以相同的信息为基础，为什么外观差别如此之大呢？—— 它们以大相径庭的形式演绎同一信息。





两张图都以同一信息为基础，外观却大相径庭。这是怎么回事？





动动笔

观察前一页的两张图。你觉得主要区别在哪里？为什么这两张图会让人对数据形成如此不同的第一印象？

世上没有傻问题

问： 为什么不直接观察数据？干嘛要用图形表示？

答： 有时候只看原始数据无法明白就里。数据中隐含着一些模式和趋势，仅仅观察堆积如山的数字很难把握这些模式和趋势。图形是发现数据隐含模式的一种有效方法。通过图形，数据得以直观地体现，使你一眼就能看出数据的真正动向。

问： 信息与数据有何区别？

答： “数据”指的是所搜集的原始事实与数字。“信息”指的是加入了某种意义的数据。

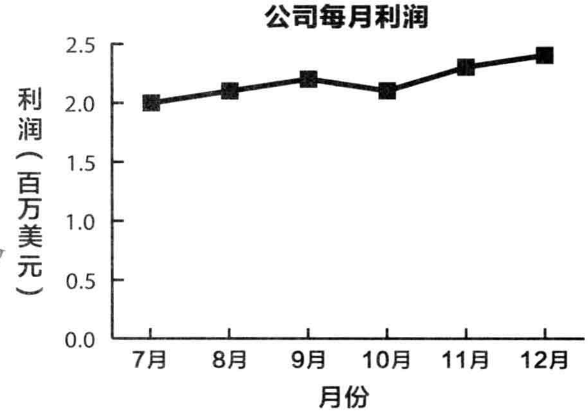
以数字5、6、7为例，单看它们本身，它们只不过是一些数字，你并不知道这些数字有何含义、代表什么——这叫做数据；随后，如果有人告诉你，这是三个孩子的年龄，你就拥有了信息，因为这些数字现在有意义了。

动动笔解答

观察这两张图，你觉得主要区别在哪里？为什么这两张图会让人对数据形成如此不同的第一印象？

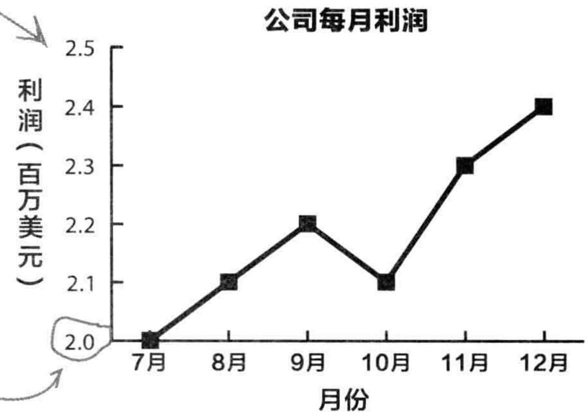
构成两张图的数据基础相同，却传递出不同的信息。

第一张图表明，利润相对稳定。之所以这样，是因为这张图的纵轴以0为原点，据此绘制每个月的利润。



看，两张图的纵轴不一样。

第二张图给人不同的印象，因为它的纵轴起点发生了变化，标度也相应发生了变化。乍一看，每个月的利润显得上涨显著。只有细细查看，你才会明白到底是怎么回事。



这张图的纵轴从2.0开始，而非从0开始，怪不得利润表现如此惊人。

我为什么要操心怎么画图啊？制图软件可以帮我们搞定一切，它就派这用场。



软件无法替你思考

制图软件可以为你节省大量时间，生成有效的图表，但你仍需了解事情的来龙去脉。

归根结底，这是你的数据。能否为自己的工作选择合适的图表，确保数据以最有效的方式展现出来并传达你想传达的信息，这取决于你。

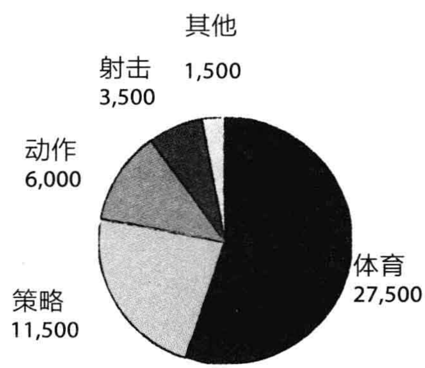
软件能够将数据转换成图表，至于图表是否正确，这得由**你**来保证。

芒芒游戏公司需要绘制图表

芒芒游戏公司

芒芒公司是一家富有创意的游戏软件公司，如今在全球市场风头正劲。公司首席执行官受邀在下届全球游戏博览会上发表主题演讲，他需要用一些巧妙、直接的方法展示数据，于是找到了你，让你给他搞出这些东西。此事关系重大，若主题演讲发表顺利，芒芒将会得到额外赞助，而你呢，肯定会因为工作努力到手大笔奖金。

首席执行官希望能够办到的第一件事是对各种游戏的满意玩家百分数进行比较。他已经动手用一些绘图软件处理过手头的数据，结果如下：



各种游戏销量

动动脑

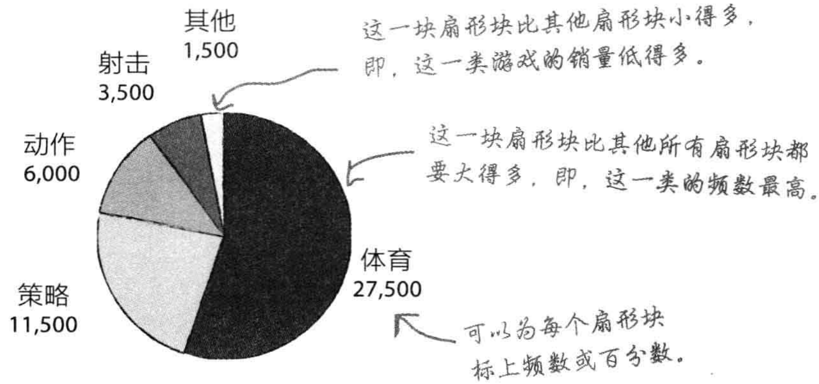
好好看看首席执行官生成的饼图。每一个小块代表什么？猜猜看，各种视频游戏的相对受欢迎程度如何？

呆板的饼图

“饼图”的作用是将数据划分为互有明显区别的几个组，或者叫做几个类。饼图为圆形，被分割为几个扇形块，每一块代表一个组（类）。扇形块的大小表示这类数据占总体的比例。扇形块越大，该组（类）的相对频繁程度越大。一个特定组中的对象数目称为**频数**。

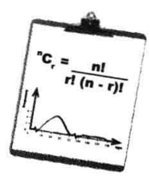
饼图将整个数据集划分为几个互不相干的组。这意味着，如果把每个扇形块的频数加起来，结果应为100%。

让我们好好看看体现了各种游戏软件销量的饼图：



各种游戏销量

种类	销量（件）
体育	27,500
策略	11,500
动作	6,000
射击	3,500
其他	1,500



重要统计量

频数

“频数”表示在一个特定组，或者说在一个特定区间内的统计对象的数目，类似于数数。

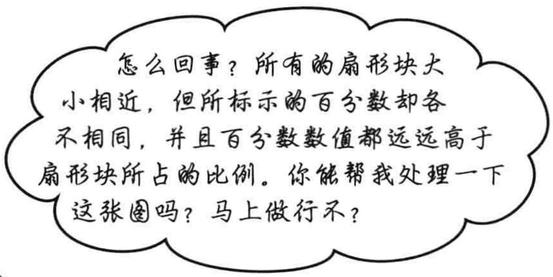
那么，饼图什么时候有用？

前面讲过，每个扇形块的大小代表你所展示的每组数据的相对频数。因此，在想对基本比例进行比较的时候，饼图有用。通过与其他组进行比较，通常很容易一眼看出哪个组具有较高频数。当所有扇形块的大小相似时，饼图用处不大，因为这时难以根据扇形块尺寸上的微小差别进行判别。

那么，芒芒首席执行官创建的饼图有用吗？

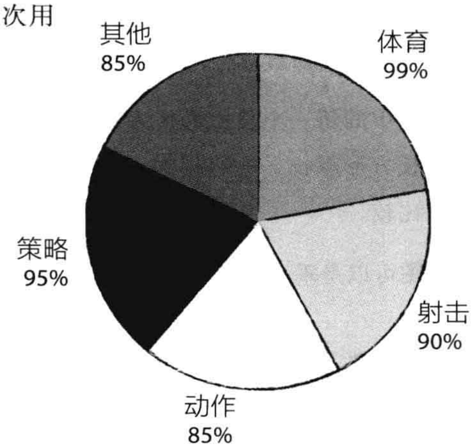
图形遇挫

看到创建一张饼图能如此出色地体现每种游戏的销量，于是，首席执行官决定再创建一张图，用以展现消费者对芒芒游戏的满意度。首席执行官需要这样一张图：能让他对每种游戏的满意玩家百分数进行比较。他再次用制图软件倒腾了一下数据，但是这一次，他感觉并不好。





芒芒首席执行官



每种游戏的满意玩家(%)

饼图的作用是对不同组（或者类）所占的比例进行比较，但在这个例子中，各个组的比例相差无几。

很难一眼看出哪一类玩家的满意度最高。

用与扇形块所占整体比例无关的百分数来标识饼图通常也会让人犯晕，例如，“体育”块标示着99%，但这一块在饼图中所占的比例仅为20%左右。另一个问题是，我们不知道每种游戏的反馈数目是否相等，因此也无法知道用这种方式对满意度进行比较是否公正。

饼图体现比例

动动脑

看一看数据，想一想这张图有什么问题。对于这种信息，用哪种图来表现更好？

条形图更具精确性

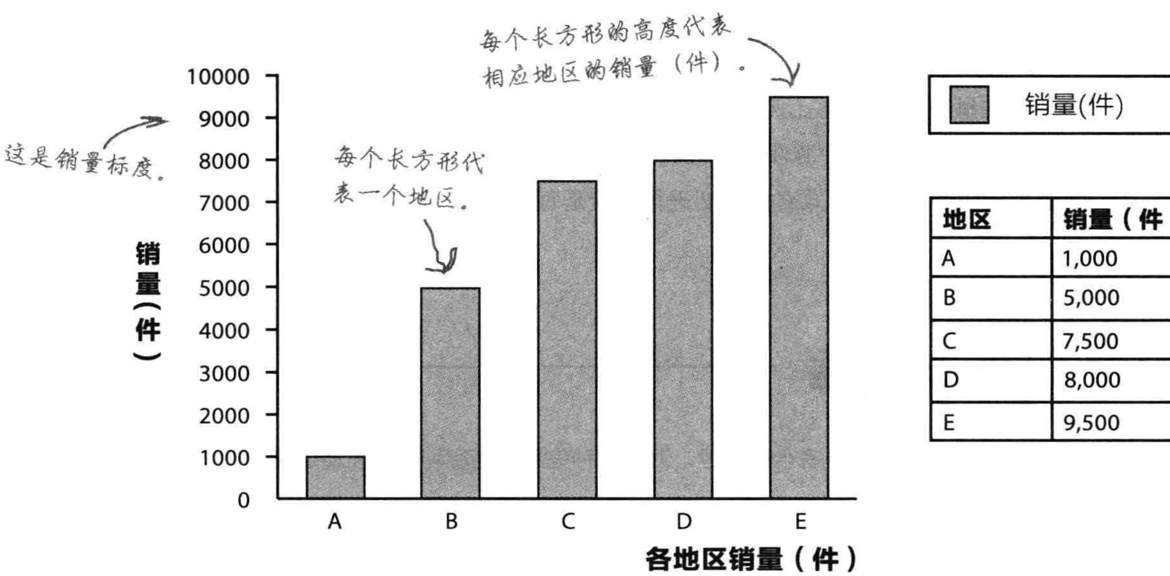
体现这种数据的更好办法是使用**条形图**。就像饼图一样，条形图能让你对相对大小进行比较，但条形图还有这样一个优点：更精确。对于各个类的大小大致相同的情况，条形图是理想的图形，你能更精确地指出哪个类的频数最高，也更容易发现细小的差别。

条形图中的每一个长方形代表一个特定类，长方形的长度代表某种数值。长方形越长，数值越大。所有长方形的宽度都相等，这样更容易进行比较。

条形图可以是垂直的，也可以是水平的。

垂直条形图

垂直条形图用横轴表示类，用纵轴表示频数或百分数。每个长方形的高度代表相应类的数值。下面这个例子体现了五个地区（A、B、C、D、E）的销量（件）。

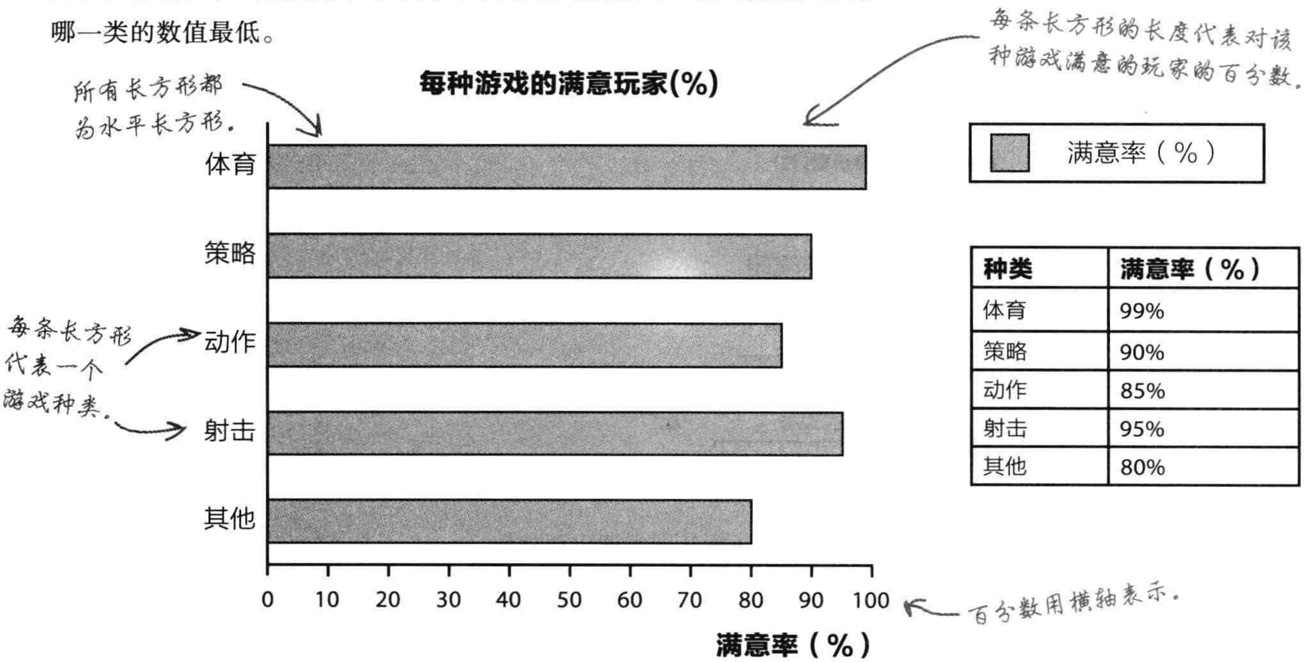


地区	销量（件）
A	1,000
B	5,000
C	7,500
D	8,000
E	9,500

水平条形图

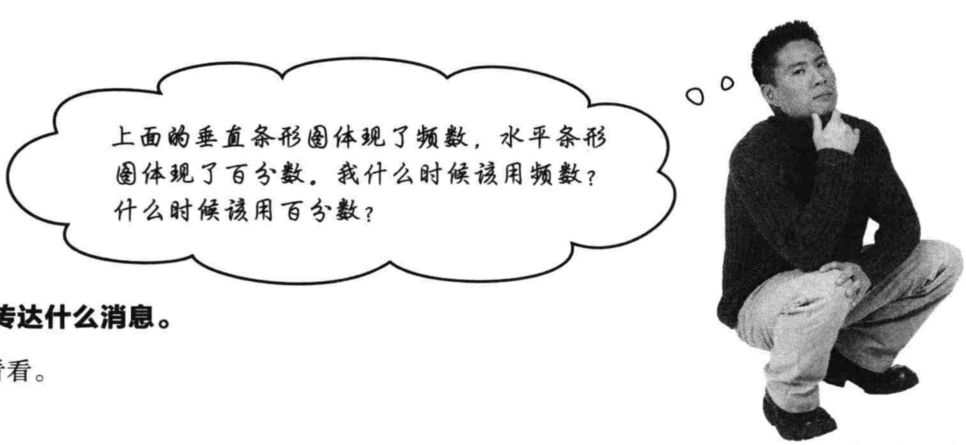
水平条形图和垂直条形图一样，只不过两根轴对调了一下。水平条形图用纵轴代表类，用横轴代表频数或百分数。

下面是用第9页上首席执行官的各类游戏满意玩家数据生成的水平条形图。如你所见，利用这张图能更快、更容易地衡量哪一类的数值最高，哪一类的数值最低。



种类	满意率（%）
体育	99%
策略	90%
动作	85%
射击	95%
其他	80%

垂直条形图更常用。不过，如果类名称太长，水平条形图就有用了——你将有大量空白位置标示每个类的名称，不用横七竖八地进行摆布。



这要看你想传达什么消息。

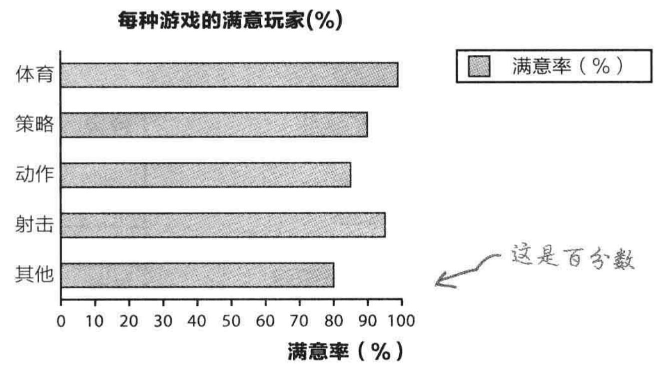
让我们好好看看。

标度的影响力

懂得利用“标度”能让你创建强大的条形图，凸显你希望别人注意的主要事实。不过，小心哦—— 标度同样能隐匿与数据有关的重要事实。下面让我们看看具体情况。

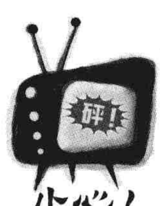
使用百分数标度

让我们先来好好看看体现每种游戏的玩家满意度的条形图。横轴表示玩家满意度百分数，即每100个人中有多少人对这款游戏感到满意。



这张图的目的是让我们对不同的百分数进行比较，还能从图中读出百分数。

只是有一个问题—— 图中没有告诉我们每种游戏有多少玩家。这听起来好像不是特别重要，但意味着我们无法知道这张图反映的是所有玩家的看法呢，还是部分玩家的看法，或甚至只是屈指可数的几个玩家的看法。换句话说，我们无法知道这能在多大程度上代表“玩家”这个整体。在设计以百分数为表现内容的图形时，请考虑这样一条黄金定律：设法指出频数——或是将频数标在图形中间，或是标在图形旁边，均可。

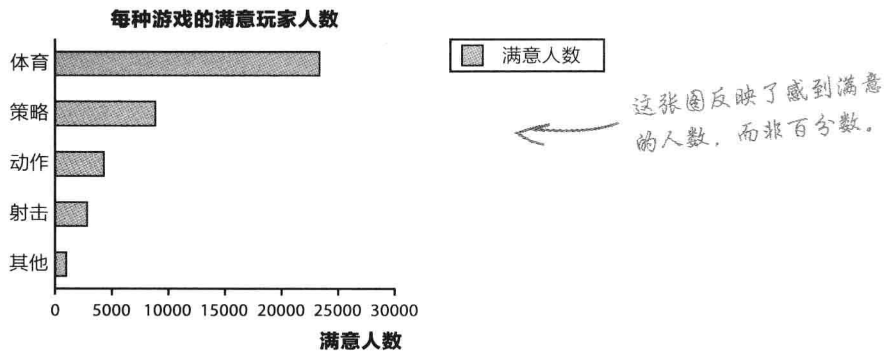


若只有百分数而没有频数，或只有频数而没有百分数，那可千万要小心。

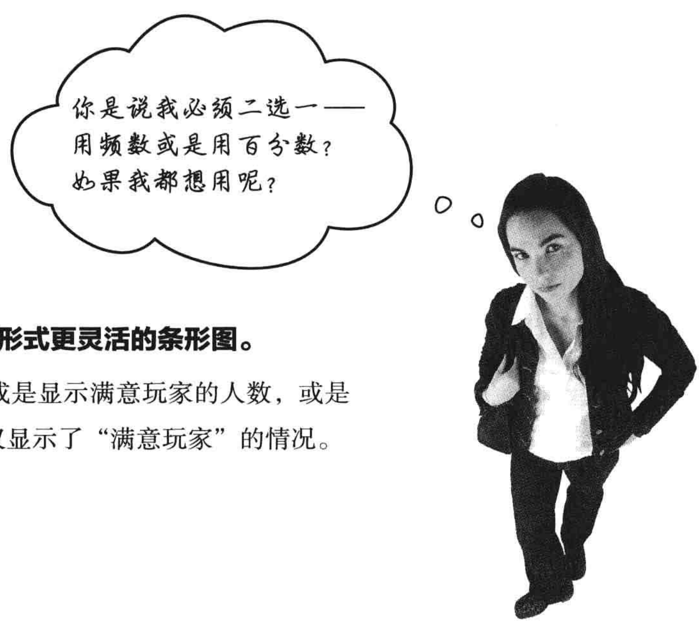
有时候，这是一种用来隐藏基础数据真实情况的伎俩，因为仅靠一张图无法判断这张图能在多大程度上代表整个数据。你可能会发现，有很大比例的人青睐某种特定游戏类别，但受到调查的仅有10人；或者，你可能会发现，有10,000个玩家最喜欢玩的是体育游戏，但仅通过这个数据无法判断这个人数在所有游戏玩家中占有的比例是高还是低。

使用频数标度

你可以用频数标度代替百分数标度。这样大家就很容易看出确切的频数，进而对数值进行比较。



通常，标度以0为起点。但要小心！并非每张图都是这么做的，正如第6页看到的，使用**不**以0为起点的标度可以让数据给人不同的第一印象，查看别人绘制的图时，要小心这一点，这很容易让你无视某些数据，从而对数据形成错误的印象。



有一些绘图方法能够绘制出表现形式更灵活的条形图。

以上这些条形图的问题是，它们或是显示满意玩家的人数，或是显示满意玩家的百分数 —— 但仅仅显示了“满意玩家”的情况。

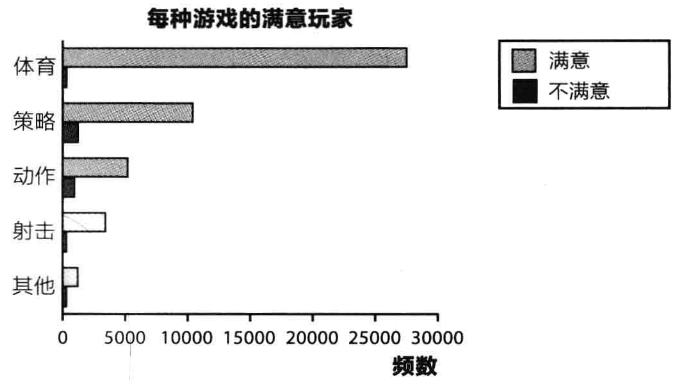
让我们看看如何解决这个问题。

处理多批数据

实际上，通过条形图能够轻而易举地在同一张图形上展现多批数据。举个例子：我们可以将满意玩家的频数和不满意玩家的频数都画在同一张图上。

堆积条形图

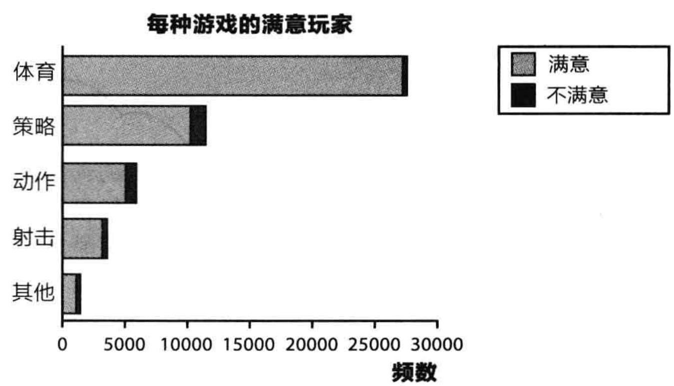
解决这个问题的一个办法是：针对每种游戏，用一条长方形代表这类游戏的满意玩家频数，用另一条长方形代表这类游戏的不满意玩家频数。当你想**比较频数时**，这种图很有用，但通过这张图难以看出比例和百分数。



分段条形图

若要同时体现频数**和**百分数，可以试试“分段条形图”。这种图用一整段长方形代表一个类，但可以按比例把这一整段长方形分割成几小段。长方形的整体长度反映出整体频数。

通过这种图可以迅速看出每个类的总频数——在我们的例子中即每种游戏的玩家总数；可以看出满意玩家的频数；还可以一眼看出比例。





练习

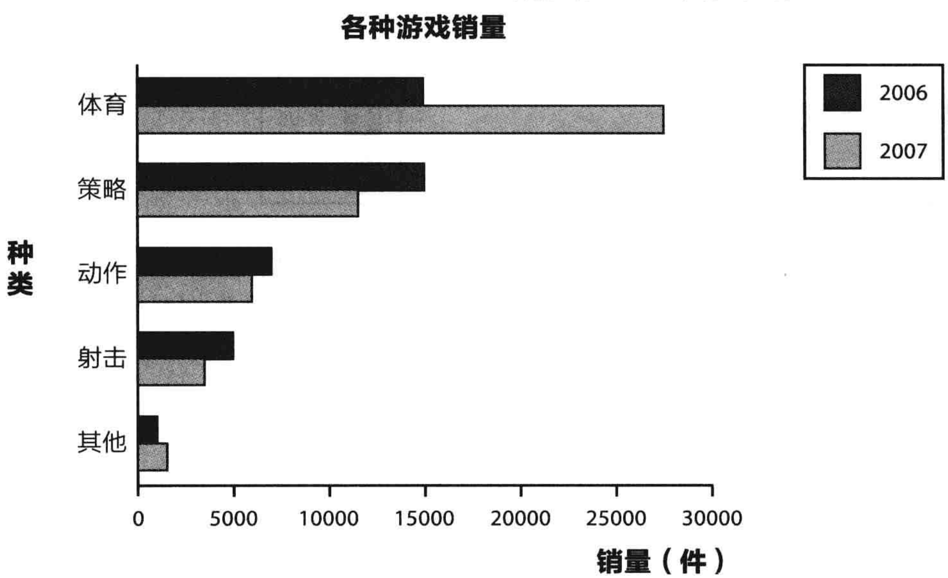
首席执行官需要为主题演讲绘制另一张图。下面是数据，看你能不能画一张条形图。

大洲	销量(件)
北美洲	1,500
南美洲	500
欧洲	1,500
亚洲	2,000
大洋洲	1,000
非洲	500
南极洲	1



动动笔

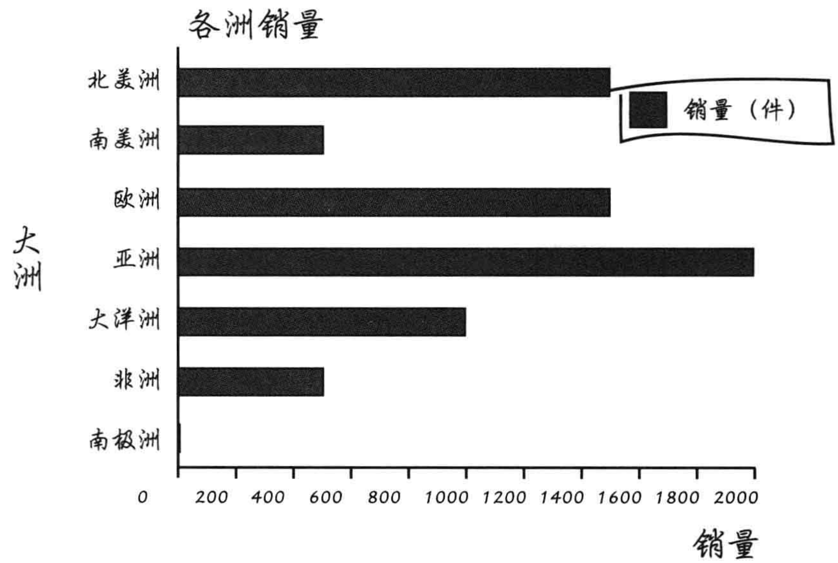
这是软件生成的另一张图。显示2007年哪种游戏卖得最好？这种游戏在2006年销量如何？



练习解答

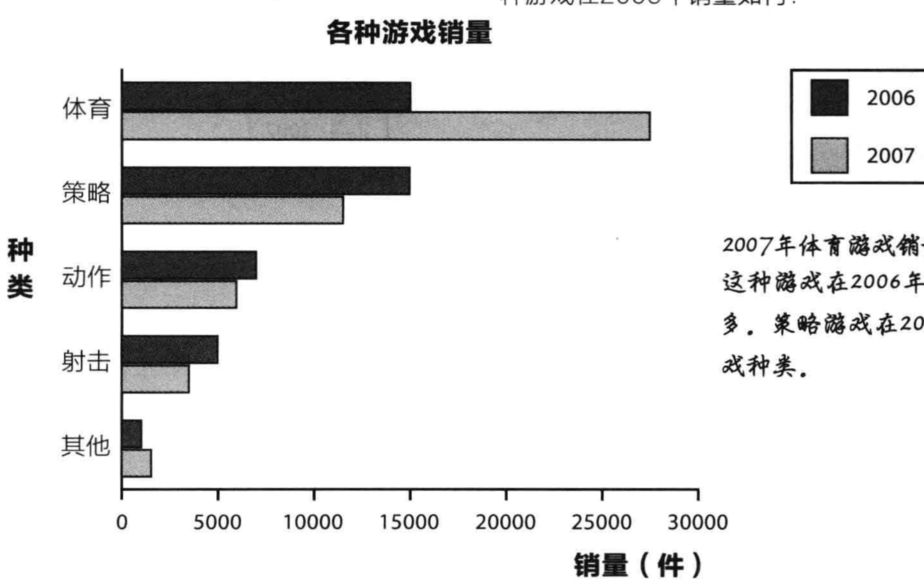
首席执行官需要为主题演讲绘制另一张图。下面是数据，看看你能不能创建这张图形。

大洲	销量(件)
北美洲	1,500
南美洲	500
欧洲	1,500
亚洲	2,000
大洋洲	1,000
非洲	500
南极洲	1



动动笔解答

这是软件生成的另一张图。显示2007年哪种游戏卖得最好？这种游戏在2006年销量如何？



2007年体育游戏销量最好，售出27,500件。这种游戏在2006年只售出14,000件，并不多。策略游戏在2006年的销量高于其他游戏种类。

你的条形图闪亮登场

首席执行官对你画出的条形图赞赏不已——但他还需要在主题演讲中报告更多数据。

干得好！这些图会在博览会上闪闪发光。现在再给你一个任务，我们请一群志愿者对新游戏进行了测试，需要用一张图来展现每局游戏的得分情况。数据如下：



游戏得分在0～999之间，得分数据被分成几个组。例如，得分在0～199范围内的次数为5。

得分	频数
0-199	5
200-399	29
400-599	56
600-799	17
800-999	3

频数为得分在某个范围中出现的次数。



这些数据看上去不同于我们之前看到过的其他类型的数据。这是不是说我们要用不同的办法进行处理？

动动脑

请回头浏览本章内容。你觉得这些数据和前面的比有什么不同吗？你觉得这种不同会对图形产生什么影响？

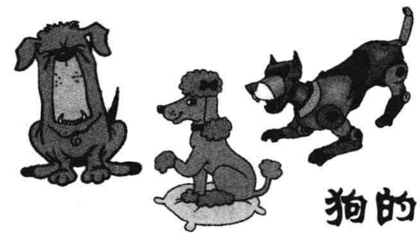
类别与数字

使用图形时，其中一个重要事项是弄清楚所处理的是哪一类数据。只要搞清楚这一点，你就会更容易决定哪一种图表能够最好地体现你的数据。

类别数据（定性数据）

目前我们讲过的大部分数据都是**类别数据**。数据被划分为各种类别，用以描述某类的性质或特征。因此，类别数据也称为**定性数据**。游戏种类就是定性数据的一个实例 —— 每个游戏种类形成一个独立的类别。

关于定性数据，请记住一个重点：不能将数据值理解为数字。



狗的品种



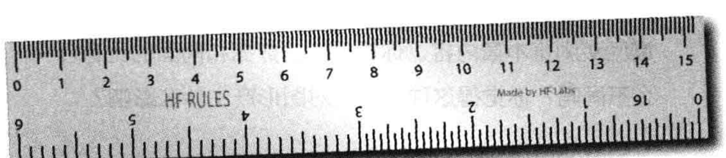
甜品的种类

数值型数据（定量数据）

数值型数据不同，它所涉及的是**数字**。数值型数据中的数值具有数字的意义，但还涉及计量或计数。由于数值型数据描述的是数量，所以也称为**定量数据**。



重量



长度



时间

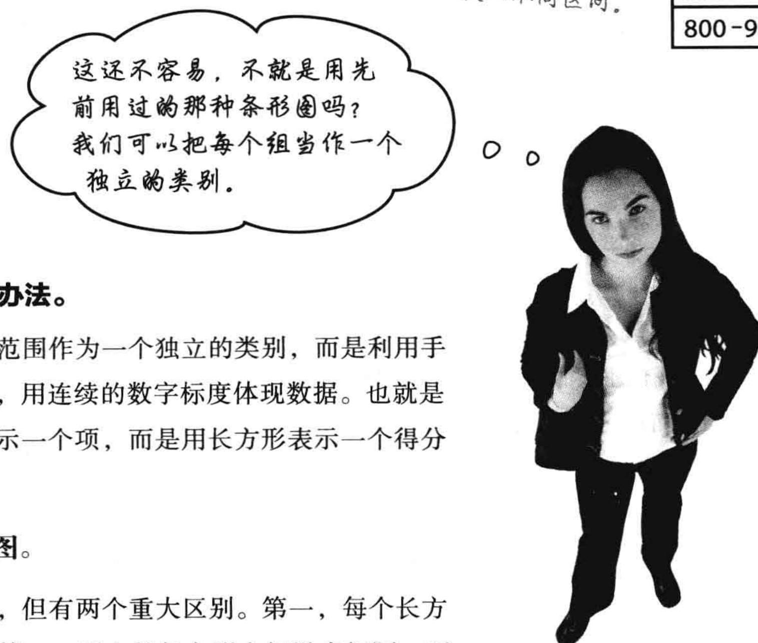
这对芒芒的图形有什么影响呢？

处理分组数据

芒芒首席执行官给我们的最新数据是数值型数据，另外，这些得分被分为几个组，放入不同的区间。那么，最好用哪种办法为这类数据绘制图形？

得分均为数字，并被分为几个组，放入不同区间。

得分	频数
0-199	5
200-399	29
400-599	56
600-799	17
800-999	3

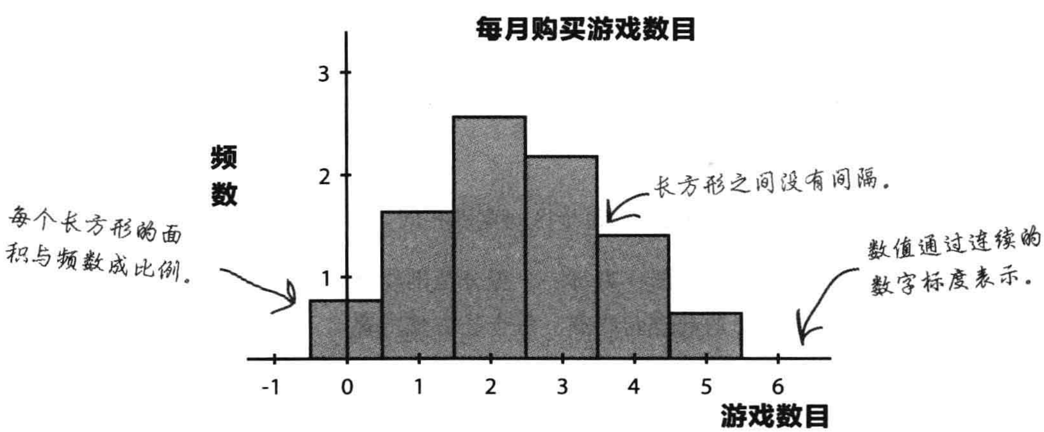


可是可以，但还有更好的办法。

我们可以不把每一个得分范围作为一个独立的类别，而是利用手头数据是数值型这一特点，用连续的数字标度体现数据。也就是说，我们不是用长方形表示一个项，而是用长方形表示一个得分**范围**。

为此，我们可以创建**直方图**。

直方图与条形图外观相似，但有两个重大区别。第一，每个长方形的面积与频数成比例；第二，图上的长方形之间没有间隔。下面是一个直方图实例，显示了统计邦中的每户人家每月购买游戏的平均数目。

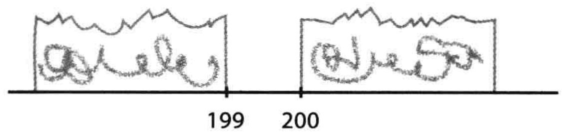


绘制直方图起步：求出长方形宽度

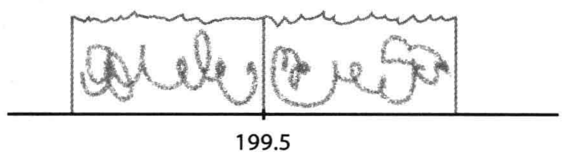
得分	频数
0–199	5
200–399	29
400–599	56
600–799	17
800–999	3

创建直方图第一步：查看每个区间，求出每个区间的宽度，以及每个区间涵盖的数据范围。同时，要确保直方图的各个长方形之间没有间隔。

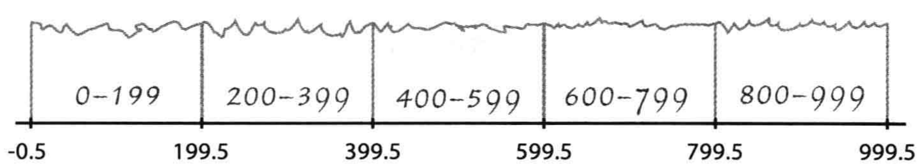
让我们从前两个区间开始：0－199和200－399。从表面数值上看，第一个区间的终点是199分，第二个区间的起点是200分。不过，要是这样画图的话，问题就来了：199和200之间将出现间隔，如下所示：



直方图的长方形之间不该有间隔。因此，为了解决上述问题，我们把以上范围稍微扩大一点儿。我们不要让第一个区间在199结束，也不要让下一个区间从200开始，而是让两个区间在199.5会合，如下所示：



这样就形成了一条唯一边界，确保直方图的长方形之间没有间隔。依法炮制其余区间，可得到下列边界：



每个区间涵盖200个得分；每个区间的宽度为200；每个区间宽度相同。

由于所有区间宽度相同，我们这样创建直方图：为每一个得分范围绘制垂直长方形，使用边界作为每个长方形的起点和终点。每个长方形的高度等于频数。

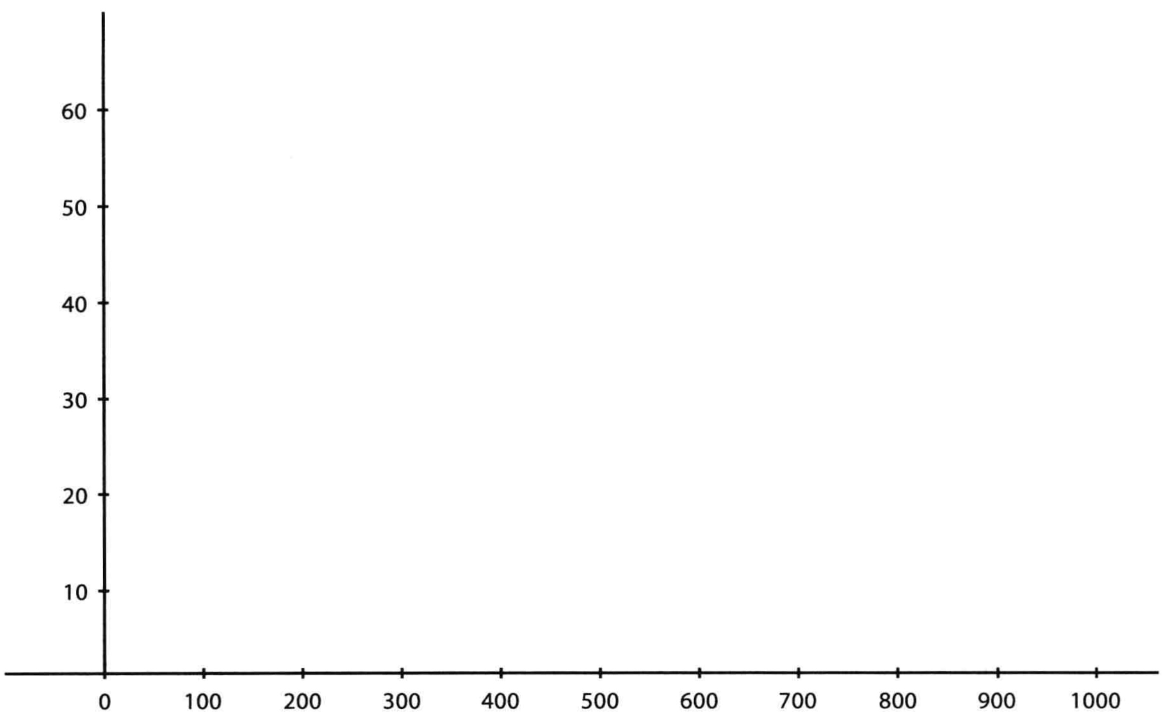


练习

下面是芒芒公司的数据备忘表。

得分	频数
0-199	5
200-399	29
400-599	56
600-799	17
800-999	3

看看你是否能利用这些边界为以上数据创建一张直方图。记住，频数位于纵轴上。



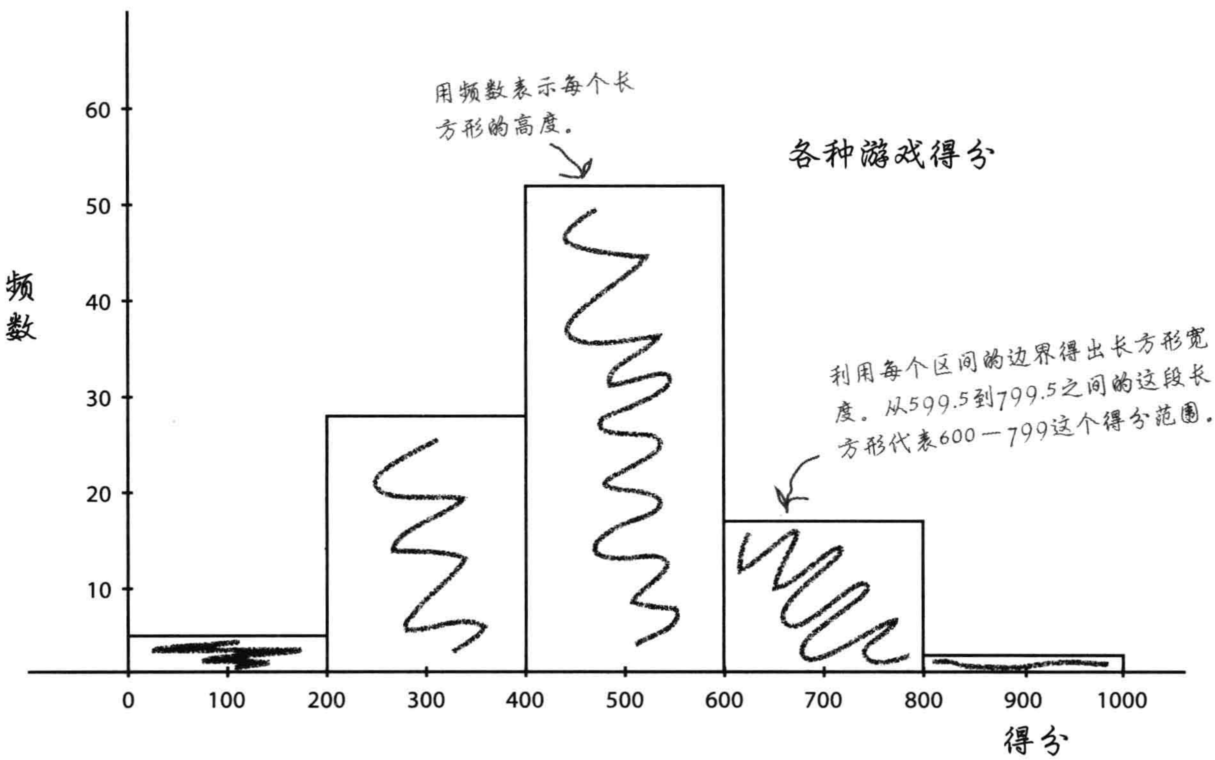


练习解答

下面是芒芒公司的数据备忘表。

得分	频数
0–199	5
200–399	29
400–599	56
600–799	17
800–999	3

看看你是否能利用这些边界为以上数据创建一张直方图。记住，频数位于纵轴上。



世上没有傻问题

问： 这么说，直方图基本上是用来体现分组数值型数据的？

答： 是的。它的优点是：由于是数值型图形，所以可以体现每个区间的宽度，还可以体现频数。

问： 如果各个区间的宽度不同会怎么样？还能使用直方图吗？

答： 完全可以。区间宽度相同是较为常见的情况，但直方图上的区间并不是非相等不可。对于区间不等的直方图，创建步骤要多两个——我们很快会介绍创建方法。

问： 直方图的长方形之间为什么不能有间隔？

答： 至少有两个有力的理由。第一是为了体现出数值之间没有间隔，每个数值都包含在内；第二是让区间宽度反映出所涵盖的数值的范围。例如，要是我们从0到199画出0－199这个区间，图上的宽度就是199 - 0=199。

问： 我们为什么要让长方形在两个数值的中间会合呢？

答： 长方形必须会合，而且通常在中间位置会合，但最终取决于所采用的舍入方法。在取整时，你通常会取离数值最近的整数，这就是说，从-0.5到0.5这个范围内的所有数值都会取整为0，于是，当我们在直方图上表示0时，我们就用从-0.5到0.5这个范围来表示0这个数。

问： 有例外吗？

答： 有，年龄就是个例外。如果你要在直方图上表示18－19这个年龄范围，通常会用18－20这个区间来表示。原因是，以19岁为例，在某人过20岁生日之前，我们通常会把他归入19岁。所以，我们用了向下取整。

要点

- **频数**是一种统计方法，用于描述一个类别中有多少个项。
- **饼图**能很好地体现基本比例。
- **条形图**更灵活、更精确。
- **数值型数据**涉及的是数字和数量；**类别数据**涉及的是表述和质量。
- 水平条形图用于展现类别数据，尤其是在类别名称太长的时候。
- **垂直条形图用于展现数值型数据**；若类别名称不长，也用于体现类别数据。
- 可以在**一张条形图上体现多批数据**，具体做法可由你选择。可以使用堆积条形图，让相互关联的长方形并列显示，借此比较频数；可以使用分段条形图，把长方形一个一个衔接起来，借此显示比例和总频数。
- **条形图标度**可以是百分数，也可以是频数。
- 每张图都变化多端。

芒芒游戏公司需要另画一图

首席执行官对你为他创建的直方图很是喜欢，所以，他想要你为他另外创建一张直方图。这一次，他想让直方图显示芒芒玩家在24小时内通常有多长时间在玩网络游戏。下面是数据：

这是玩家玩游戏的小时数

小时	频数
0–1	4,300
1–3	6,900
3–5	4,900
5–10	2,000
10–24	2,100

这是玩这么长时间游戏的玩家频数



他说对了，区间宽度并不都相同。

只要看看这些区间，就能看出它们具有不同的宽度。例如，10－24这个范围涵盖的小时数远多于0－1这个范围。

如果我们有办法得到原始数据，就可以看看如何设法构建等宽区间，但遗憾的是，我们所拥有的全部数据都在这儿了。我们需要找到这样一种绘制直方图的方法：容许数据区间具有不同宽度。

动动脑

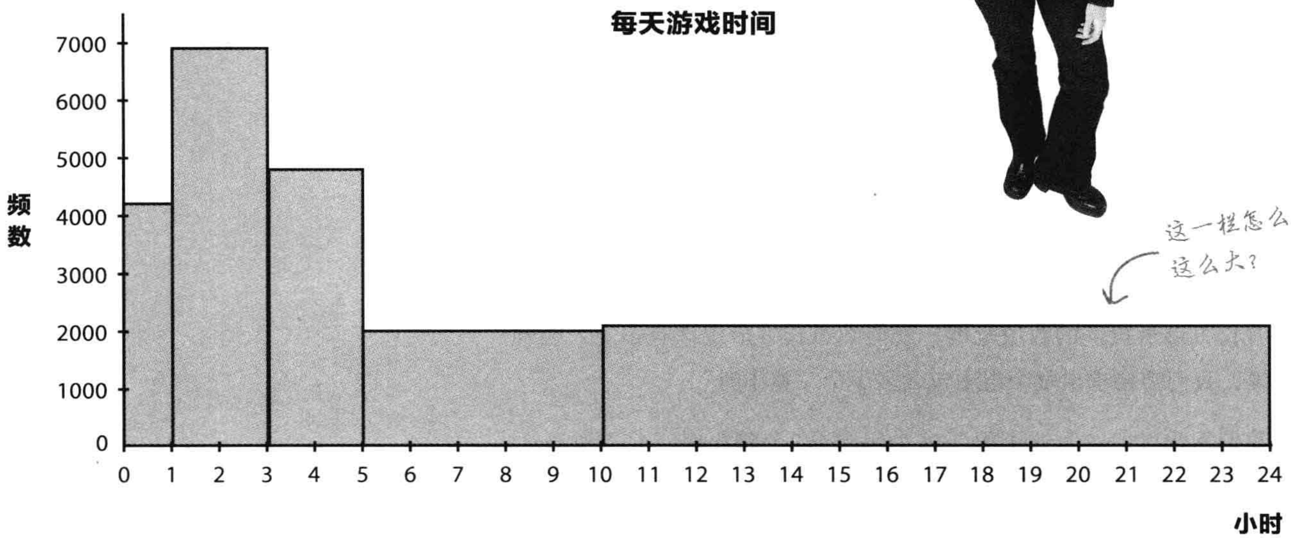
直方图的特点是：频数与每个长方形的面积成比例。你会如何利用这一点为以上数据创建直方图？你需要知道些什么？

我想我们可以用以前用过的方法创建这张图，没什么大不了的。利用数字标度画出长方形，只不过这次的长方形宽度不一样。



你认为她对吗？

下面是一张草图，垂直标度为频数，长方形宽度与区间大小成比例绘制。你看出问题了吗？



直方图的长方形面积必须与频数成比例

这张图的问题是，为了让每个长方形的宽度反映出每个区间的宽度，结果造成一些长方形看起来超大，比例失衡。乍一看，你可能对人们每天玩游戏的实际时间心生误会。例如，面积最大的长方形是显示玩游戏时间在10−24小时之间的长方形，但大部分人并不玩这么长时间。

由于这是一张直方图，我们需要让长方形面积与长方形所代表的频数成比例。长方形的宽度不相同，我们该怎么处理长方形的高度呢？

让直方图长方形的面积与频数成比例

到目前为止，我们已经能用长方形的高度表示特定数字或类别的频数了。

这一次，我们要处理分组数值型数据，这些数据的区间宽度各不相等。我们当然可以让每个长方形的宽度反映每个区间的宽度，可是这种做法的问题是：长方形具有不同宽度，这会影响每条长方形的总面积。

我们需要确保每条长方形的面积与频数成比例。这意味着，只要我们调整长方形宽度，就要同时调整长方形高度。如此一来，就能在改变长方形宽度 —— 最终使其反映分组宽度的同时，保持长方形的面积与频数相吻合。

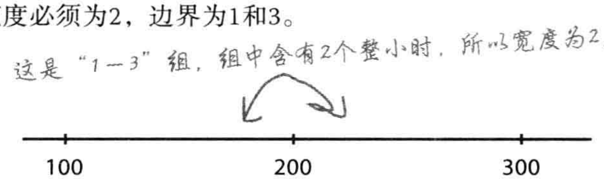
直方图的特点是：长方形面积表示频数。

让我们看看如何创建一张新直方图。

第1步：求长方形宽度

看看长方形所覆盖的数值范围，就能知道长方形应该有多宽。换句话说，我们需要求出每个组中包含多少个“整小时”。

让我们取出“1－3”这个组。这个组包含2个整小时：1－2和2－3。这表示长方形的宽度必须为2，边界为1和3。



算一算其余宽度，得出：

小时	频数	宽度
0–1	4,300	1
1–3	6,900	2
3–5	4,900	2
5–10	2,000	5
10–24	2,100	14

算出长方形宽度后，就可以接着求高度了。

第2步：求长方形高度

求出所有组的宽度后，就可以利用这些宽度求出长方形应该有的高度。别忘了，我们需要调整长方形高度，使得每个长方形的整体面积与相应组的频数成比例。

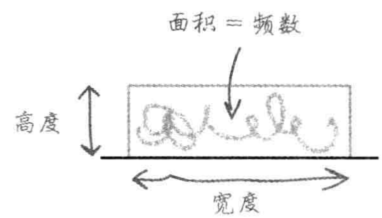
首先，让我们定下每个长方形的面积。前面说过，频数等于面积。由于我们已知每个组的频数，也就知道面积应该是多少：

长方形面积 = 每组频数

这些频数我们一开始就知道，于是我们知道目标面积是多少。

现在每个长方形基本上就是一个矩形，这意味着每个长方形的面积等于宽度乘以高度。由于面积等于频数，即：

频数 = 长方形宽度 × 长方形高度



我们在上一步求出了长方形的宽度，于是，可以用这些宽度求出每个长方形的高度。即：

$$\text{长方形高度} = \frac{\text{频数}}{\text{长方形宽度}}$$

长方形高度用于量度一个特定组的频数的**集中程度**，是对频数密集度的一种量度，是用于说明数字到底是“稠密”还是“稀薄”的一种方法。长方形的高度称为**频数密度**。

动动笔

每个长方形的高度应该是多少？填写下列表格。

小时	频数	宽度	高度（频数密度）
0-1	4,300	1	4,300 ÷ 1 = 4,300
1-3	6,900	2	
3-5	4,900	2	
5-10	2,000	5	
10-24	2,100	14	

动动笔解答

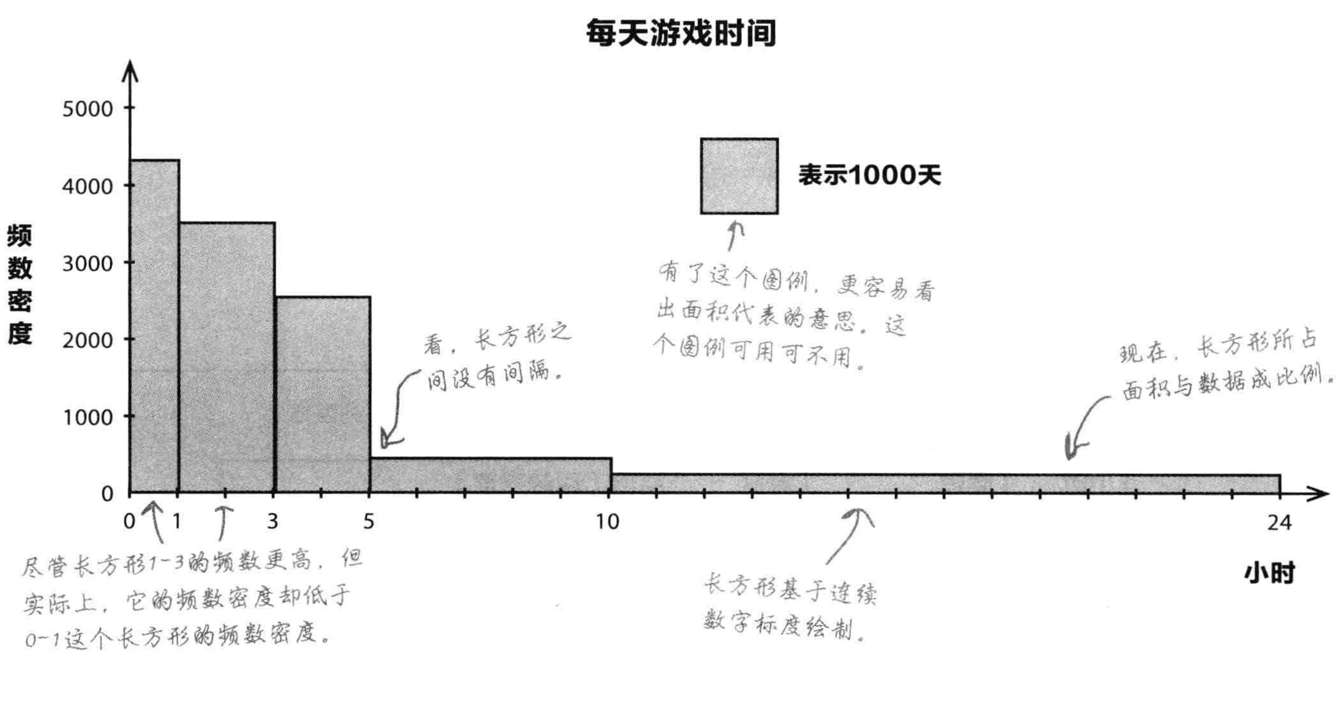
每个长方形的高度应该是多少？填写下列表格。

小时	频数	宽度	高度（频数密度）
0-1	4,300	1	4,300 ÷ 1 = 4,300
1-3	6,900	2	6,900 ÷ 2 = 3,450
3-5	4,900	2	4,900 ÷ 2 = 2,450
5-10	2,000	5	2,000 ÷ 5 = 400
10-24	2100	14	2,100 ÷ 14 = 150

第3步：画出直方图

求出每个长方形的宽度和高度之后，就能画出直方图了。画图方法和以前一样，但这次，我们为纵轴标上频数密度，而非频数。

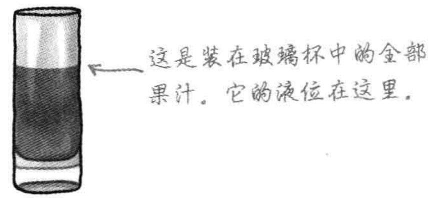
下面是经过修订的直方图。



频数密度细细看

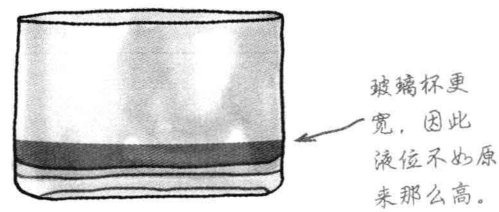
频数密度指的是数据中的数值密集度。频数密度与频数有关，但并非同一事物。下面用一个比喻来说明二者之间的关系。

想像一下，你有一些果汁，并将这些果汁倒进玻璃杯，如图所示：



要是把相同分量的果汁倒入另一个不同尺寸玻璃杯（假定“宽”一点儿），情况如何呢？果汁液位有何变化？——图中的玻璃杯宽一点儿，因此果汁液位降低了。

果汁液位随着玻璃杯的宽度发生变化，玻璃杯越宽，液位越低；反之亦然——玻璃杯越窄，果汁液位越高。

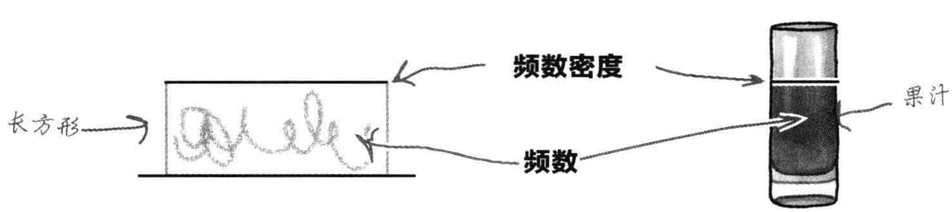


那么，果汁与频数密度有什么关系？

果汁 = 频数

这样想像：你不是在向玻璃杯中倒果汁，而是在把频数“倒入”图形中的长方形。正如你知道玻璃杯的宽度一样，你也知道长方形的宽度；正如果汁在玻璃杯中占有的空间（底面积 × 高）等于玻璃杯中的果汁的分量，图中的长方形的面积等于其频数。

这样一来，频数密度就等于长方形的高度，接着使用上面的比喻，这个高度就等于果汁在每个玻璃杯中的液位。较宽的玻璃杯意味着果汁会达到一个较低的液位，而较宽的长方形意味着频数密度会较低。



要 点

- **频数密度**指的是分组数据中的频数的密集度。计算方法如下：

$$频数密度=\frac{频数}{组距}$$

- **直方图**是一种专门用于体现分组数据的图形。它看起来很像条形图，但每条长方形的高度等于频数密度——而不是频数。
- 绘制直方图时，每个长方形的宽度与其分组宽度（“组距”）成正比例。长方形按照连续的数字标度绘制。
- 直方图中的每个组的频数通过长方形面积求出。
- 直方图的长方形之间没有间隔。

世上没有傻问题

问： 画直方图时，为什么用面积代表频数？

答： 这样做可以保证每个组的相对大小与数据成正比例，且不失真实。处理分组数据时，我们需要通过一种直观的方法体现每个组的宽度及频数。改变长方形宽度是一种反映分组范围的直觉方法，但这种方法有一个副作用——会使一些长方形看起来比例失衡。

调整长方形高度并用面积表示频数，这是解决以上问题的一个办法。有了这个办法，大家就不会由于某个组占用了太多或太少空间而产生错觉。

问： 什么又是频数密度呢？

答： 频数密度是表示某个特定区间中的数据密集度的一种方法。通过这种方法可以对宽度可能有差别的几个区间进行比较。在这种方法中，频数与长方形的面积成正比例，而不是与高度成正比例。

为了求出频数密度，应取出这个区间的频数，用它除以宽度。

问： 如果我已经将数据分组，但所有的区间都具有相同宽度，我能使用普通的条形图吗？

答： 使用直方图能更好地体现你的数据，因为你还要接着对分组数据进行处理。你确实需要让频数与面积成比例，而不是与高度成正比例。

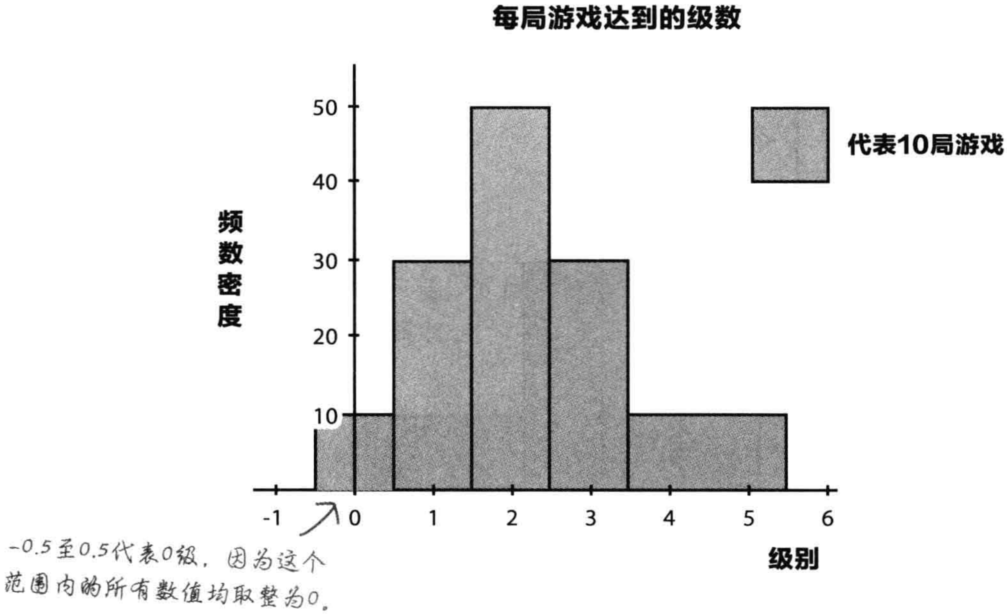
问： 直方图“必须”体现分组数据吗？能不能用于体现一个个数字及一批批数字？

答： 能。主要记住这一点：确保长方形之间没有间隔，以及每个长方形的宽度均为1。为了实现这一点，通常可将数据中的数字放在长方形的中央。

例如，如果要画一个长方形代表单独的数字1，则必须画一个范围为0.5到1.5的长方形，1位于这个范围的中央。

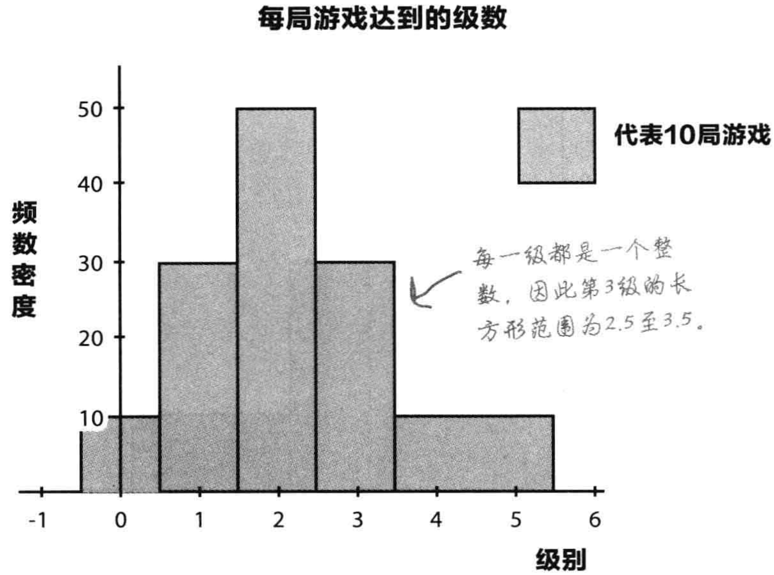


下面这张直方图体现了每打一局“疯狂奶牛”游戏达到的级数。总共打了几局游戏？假定每一级为一个整数。



练习解答

下面这张直方图体现了每打一局“疯狂奶牛”游戏达到的级数。总共打了几局游戏？假定每一级为一个整数。



我们需要求出玩游戏的总局数，也就是说，要求出总频数。

总频数等于每个长方形的面积之和。因此，我们要用每个长方形的宽度乘以该长方形的频数密度，得出频数，然后将所有频数相加。

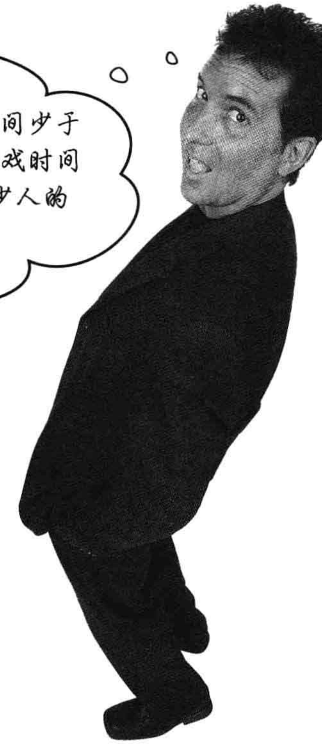
级别	宽度	频数密度	频数
0	1	10	1×10 = 10
1	1	30	1×30 = 30
2	1	50	1×50 = 50
3	1	30	1×30 = 30
4–5	2	10	2×10 = 20

总频数 = 10 + 30 + 50 + 30 + 20

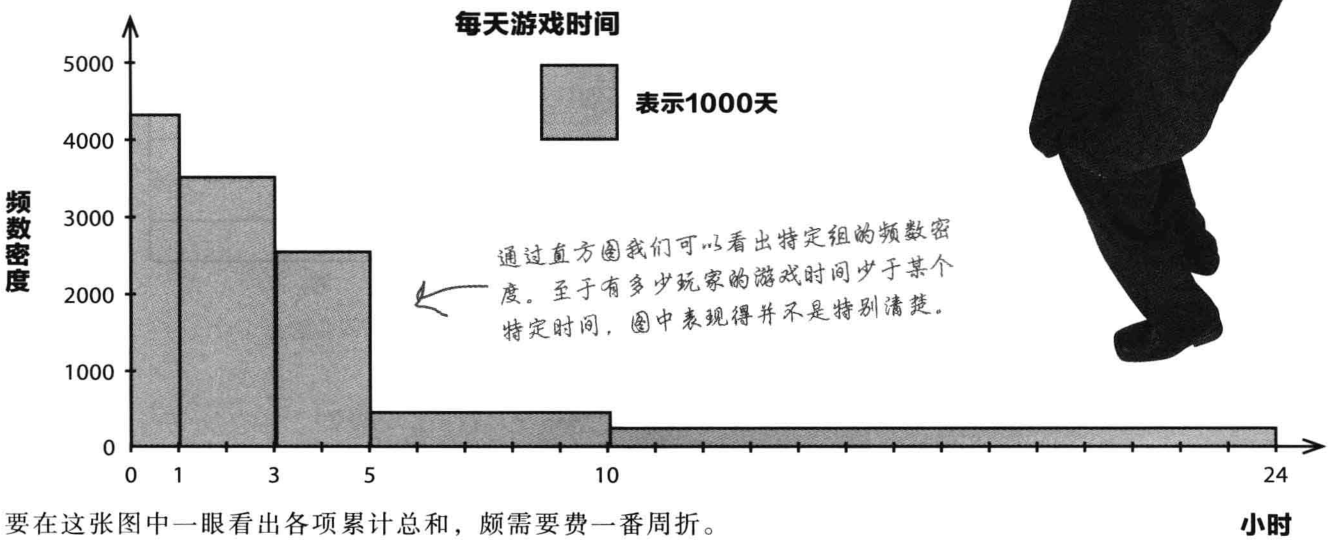
= 140

直方图并非无所不能

尽管直方图在显示分组数值型数据方面表现出色，但还有几种数值型数据用直方图表现并不理想，比如不断在原有总和上增加新值而得出的“累计总和”……



让我们看看能不能帮帮首席执行官。下面是我们曾经画过的直方图：



要在这张图中一眼看出各项累计总和，颇需要费一番周折。为了求出游戏时间在5小时以内的玩家的频数，我们需要将各种频数加起来。我们需要另一种图形……哪一种呢？

动动脑

你认为我们该在图上显示哪些信息呢？该画哪些信息？请写下答案。

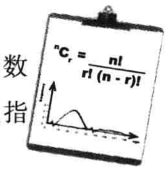
认识累积频数

首席执行官希望有某种图形能向他显示某个特定值以内的频数之和——**累积频数**。提到累积频数这个术语时，我们基本上指的是累计总和（向原来的总和中增加新值得出的总和）。

我们需要画出这样的图：用横轴表示时间（小时），用纵轴表示累积频数。通过这张图，首席执行官就能取一个值，并从图上读出到这个数值为止的相应累积频数。他将能求出游戏时间在5小时内、6小时内或他最感兴趣的任意小时内的人数。

在动手画图之前，我们需要知道到底要在图上画些什么——我们需要计算已知的每个区间的累积频数，还要求出每个区间的上限。

让我们看看数据，开工！



重要统计量

累积频数

累加到某个数值为止的总频数。基本上是所有频数的累计总和。

小时	频数
0-1	4,300
1-3	6,900
3-5	4,900
5-10	2,000
10-24	2,100

这是数据。

那么，累积频数是多少？

首先，让我们假定首席执行官需要画出1小时以内的累积频数（或者叫总频数）。只要我们看看数据就知道，0–1组的频数是4300，还能看出1是该组的上限。即，在1小时以内，累积频数为4300。

接下来，看看3以内的总频数。我们已知0–1组和1–3组的频数，3是又一个上限。为了求出3以内的总频数，我们将0–1组和1–3组内的频数加起来。

看出某种模式了吗？如果我们取每个组的上限（小时），将这个上限以内的各个频数相加，就能求出至该上限为止的总频数，以此类推，得出：

小时	频数	上限	累积频数
0	0	0	0
0-1	4,300	1	4,300
1-3	6,900	3	4,300+6,900 = 11,200
3-5	4,900	5	4,300+6,900+4,900 = 16,100
5-10	2,000	10	4,300+6,900+4,900+2,000 = 18,100
10-24	2,100	24	4,300+6,900+4,900+2,000+2,100 = 20,200

加数为0，因为每周玩游戏时间不会少于0小时。

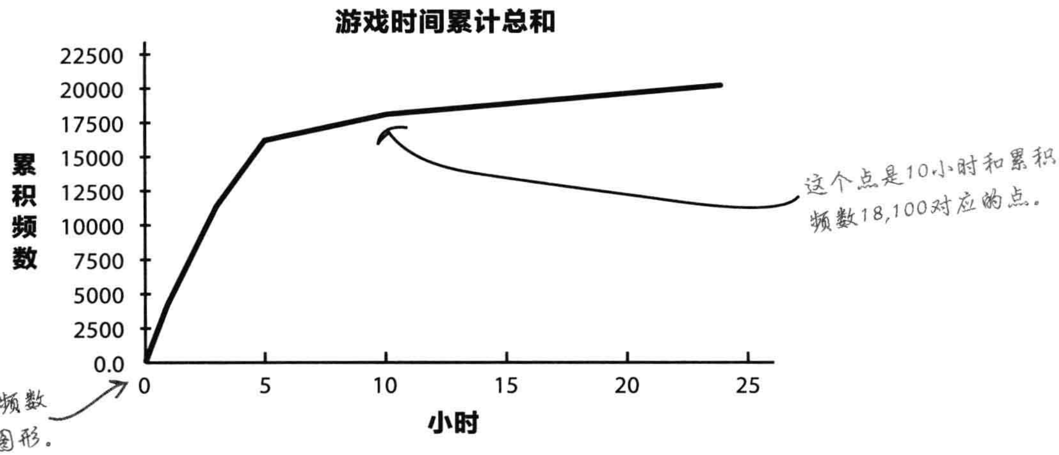
绘制累积频数图

既然已经有了各个上限和累积频数，我们就能在图上画出这些数据了。画两条轴，纵轴代表累积频数，横轴代表小时数。画好后，根据上限及与之对应的累积频数画出各个点，然后用一条线将这些点连起来，如下图：



累积频数决不会减小。

只要发现累积频数开始减小，就应检查计算方法是否正确。

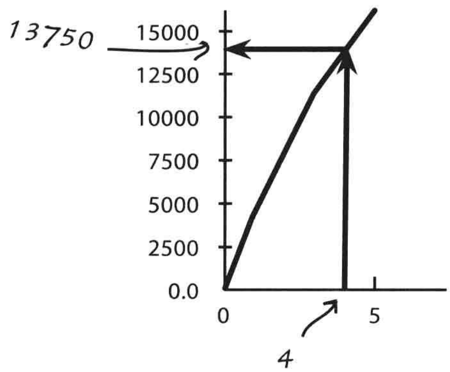


动动笔

首席执行官想让你求出人们在线游戏时间在4小时以内的发生次数。看看能不能用累积频数图估计这个值。

动动笔解答

首席执行官想让你求出人们在线游戏时间在4小时以内的发生次数。看看能不能用累积频数图估计这个值。



为此，我们在横轴上找到4，找到这个数值与图线的交点，然后读出纵轴上的相应累积频数。

由此得出答案约为13,750。换言之，在线游戏时间在4小时以内的约有13,750次。

世上没有傻问题

问： 什么是累积频数？

答： 某个数值的累积频数即到这个数值为止（包括这个数值在内）的频数总和。通过累积频数可知到该数值点为止的总频数。

例如，假设你有一些人的年龄数据。数值27的累积频数表示到27岁（包括27岁在内）为止的人有多少。

问： 累积频数只是用于分组数据吗？

答： 完全不是。累积频数可以用于任何数值型数据。关键是，你想知道的是到某个特定数值为止的总频数，还是对特定数值的频数更感兴趣。

问： 有些图形可以在一张图上显示多批数据。累积频数图行吗？

答： 可以。在累积频数图上可以这样做：为每一批数据绘制一条单独的线条。例如，如果你想按性别比较累积频数，就可以画一条线表示男性，另画一条线表示女性。将两条线画在同一张图上效果会好得多，可以更容易地比较两批数据。

问： 在同一张图上绘制的线条的数目是否受到限制？

答： 没有什么特别的限制，这完全取决于你的数据。但图上线条过多会显得拥挤，这时无法在图上读出累积频数，也无法比较各个批次的数据，因此不要画过多的线条。

问： 请提醒一下我，如何求出某个数据的累积频数？

答： 可以直接从图上读出累积频数：在横轴上找到要求其累积频数的数值，找到这个数值与累积频数曲线的交点，然后从纵轴上读出累积频数的数值。

问： 如果已知累积频数，能通过图形求出相应的数值吗？

答： 能。在纵轴上找到要求其数值的累积频数，找到这个累积频数与累积频数曲线的交点，然后读出相应横轴数值。

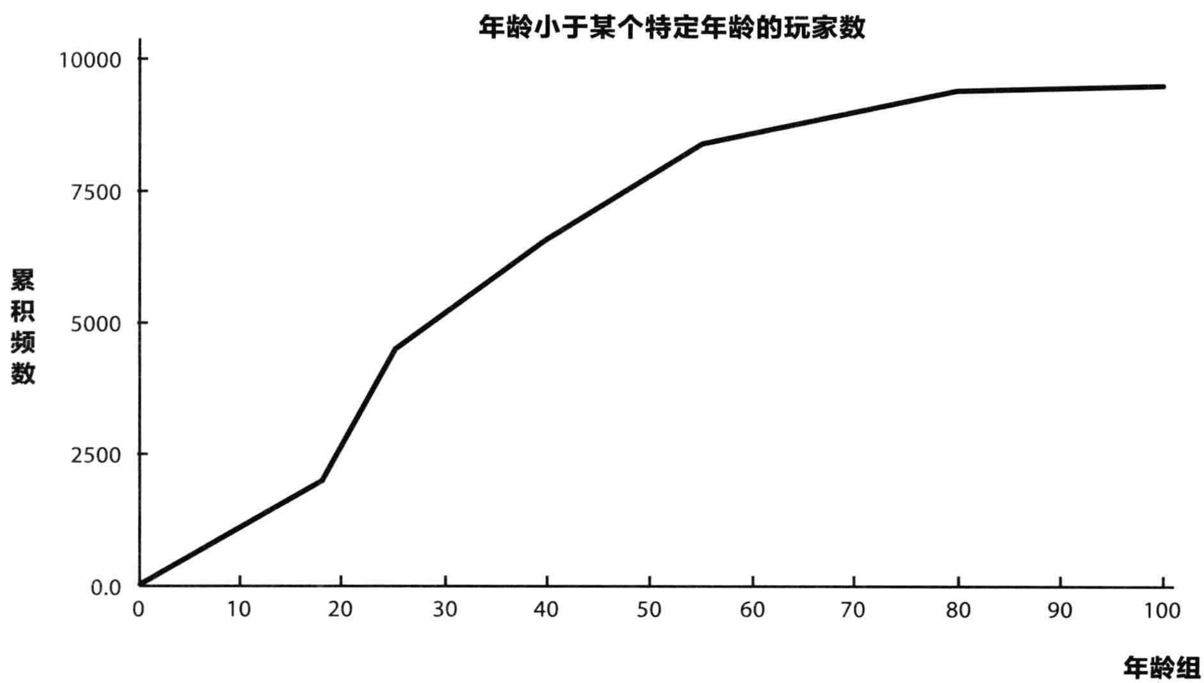


练习

在芒芒游戏公司的主题报告中，首席执行官想说明他要如何定位特定的年龄组。他有显示年龄累积频数的累积频数图，但他同时需要显示频数。可一只狗吞吃了写有这些频数的纸张。看看你是否能用累积频数图估计出每个组的频数。

这里的上限为18，因为某人从进入17岁开始到年满18岁为止均被当作17岁，年龄通常向下取整。

年龄组	上限	累积频数	频数
<0	0	0	0
0-17	18		
18-24			
25-39			
40-54			
55-79			
80-99			





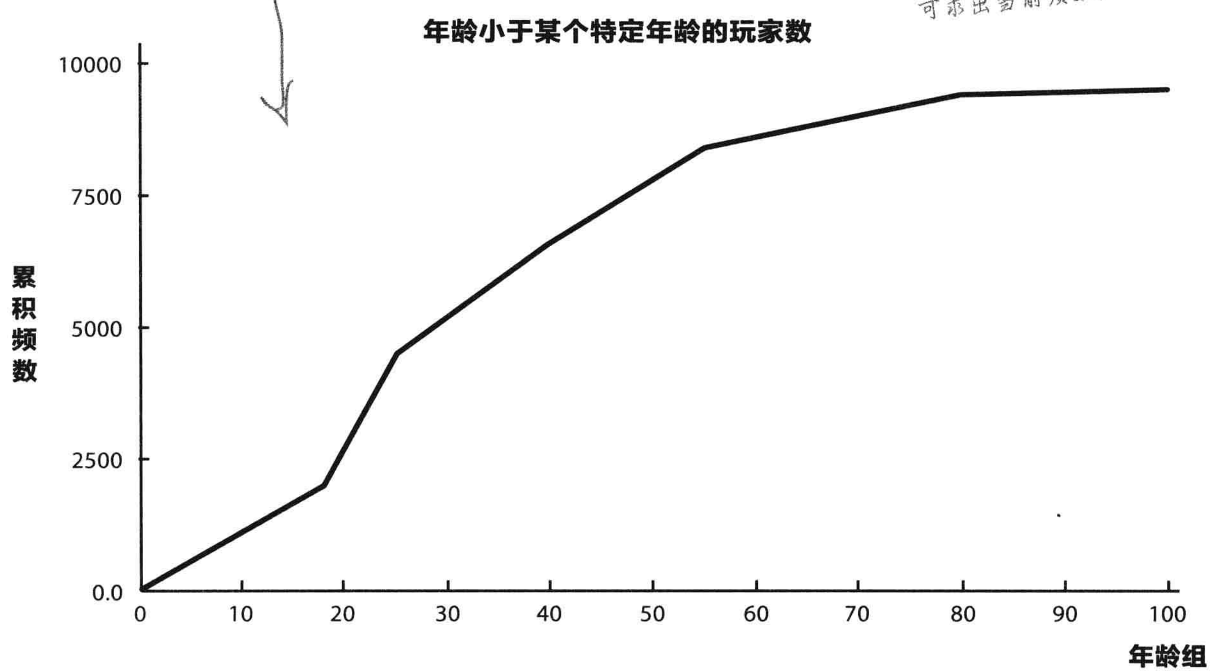
在芒芒主题报告中，首席执行官想说明他要如何定位特定的年龄组。他有显示年龄累积频数的累积频数图，但他同时需要显示频数。可一只狗吞吃了写有这些频数的纸张。看看你是否能用累积频数图估计出每个组的频数。

年龄组	上限	累积频数	频数
<0	0	0	0
0-17	18	2,000	2,000
18-24	25	4,500	4,500−2,000=2,500
25-39	40	6,500	6,500−4,500=2,000
40-54	55	8,500	8,500−6,500=2,000
55-79	80	9,400	9,400−8,500=900
80-99	100	9,500	9,500−9,400=100

在图上找出累积频数。

纵然略有误差也不用担心，这些只是估计值。

用当前累积频数减去之前的累积频数，即可求出当前频数。



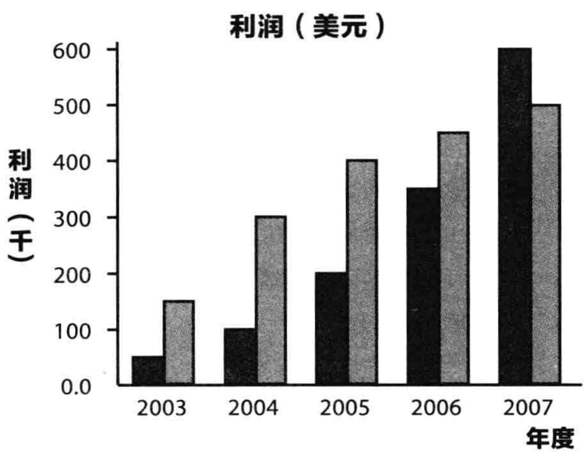
选择正确的图形

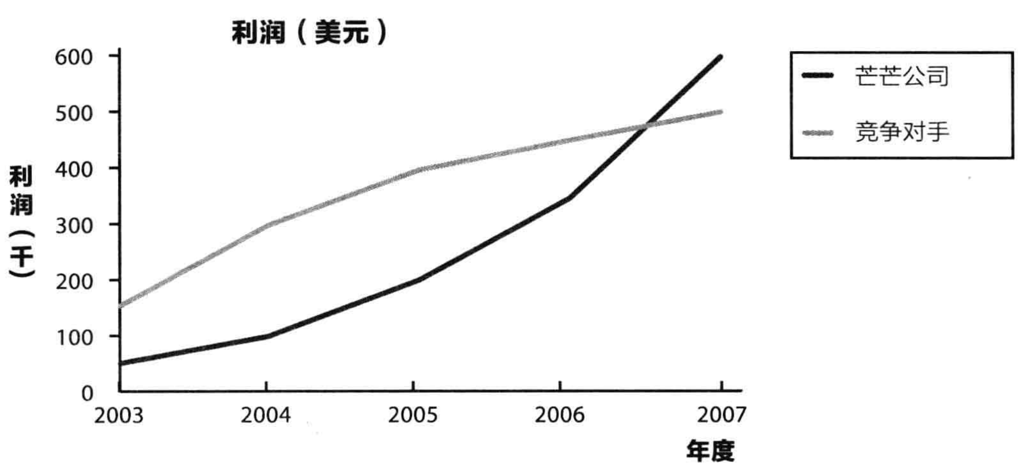
首席执行官对你绘制的累积频数图满意极了，你的奖金即将落袋为安。他已经快完成主题报告的准备工作，只差最后一图：芒芒公司与主要竞争对手利润对比图。他该用哪种图呢？



练习

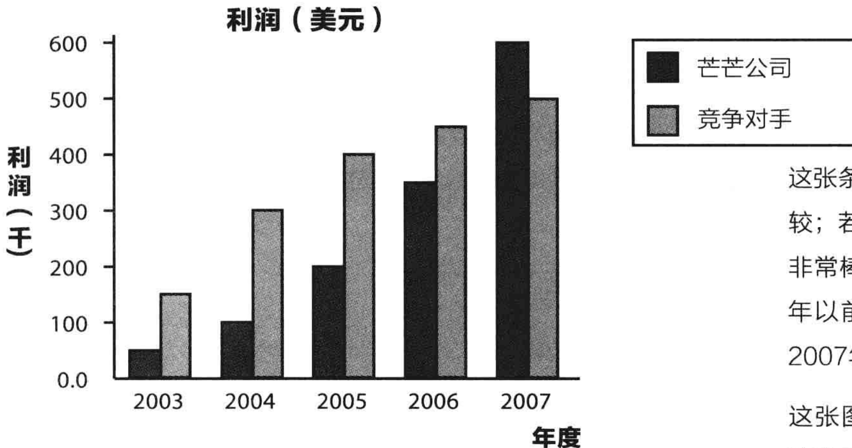
下面是首席执行官有可能用到的两张图。你的任务是辨析这两张图，对每张图的相对优缺点发表看法。你将选择哪张图？





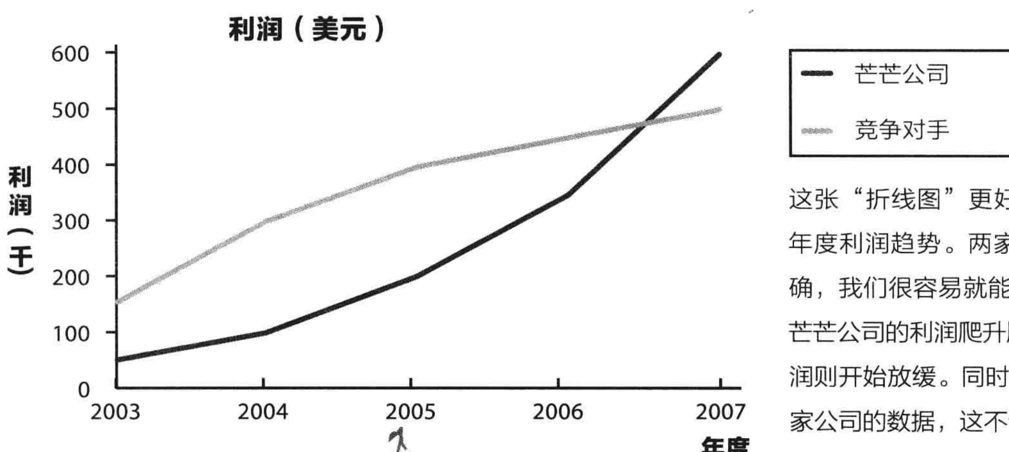
练习解答

下面是首席执行官有可能用到的两张图。你的任务是辨析这两张图，对每张图的相对优缺点发表看法。你将选择哪张图？



这张条形图按年度对利润进行了很好的比较；若想比较同一年度的利润，这张图也非常棒。例如，我们可以看出，在2007年以前，竞争对手的利润较高，但到了2007年，芒芒公司的利润超过了对手。

这张图的缺点是，如果首席执行官突然决定在图中添加第三家竞争对手的数据，读图难度可能会增加，人们难以一眼看明白这张图。



我们会选择折线图，因为它的整体趋势比条形图更清晰。不过，就算选了条形图也不用担心——选用哪种图形取决于所要强调的主要事实。

这张“折线图”更好地体现出每家公司的年度利润趋势。两家公司的趋势线都很明确，我们很容易就能看出他们的利润模式：芒芒公司的利润爬升顺利，而竞争对手的利润则开始放缓。同时，很容易就能添加另一家公司的数据，这不会让图形面目不清。

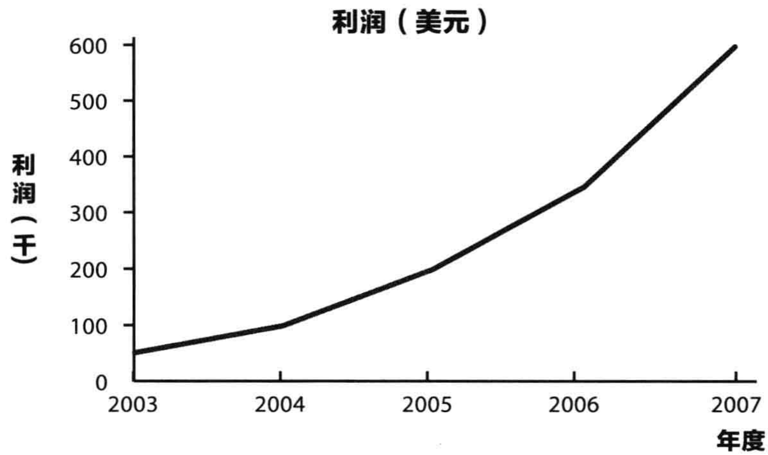
缺点是，虽然也能够对年度利润进行比较，但不如条形图清晰。

折线图细细看

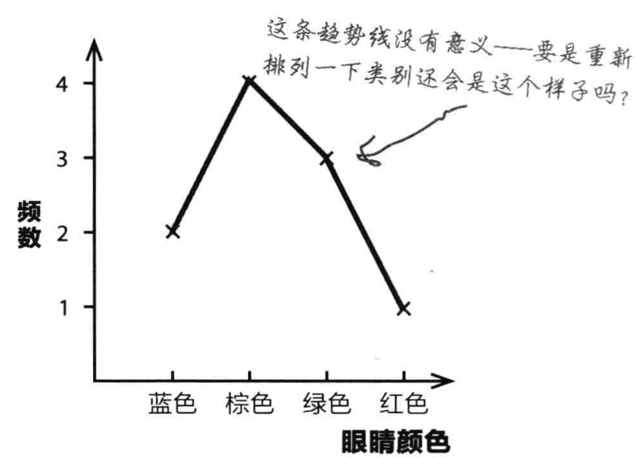
折线图能很好地体现数据趋势。你将每一批数据画成点，然后将这些点连起来。就可以方便地在同一张图上显示多批数据，却不会显得过于拥挤 —— 只要确保能清楚地看出每一条线就行了。

像其他图形一样，在纵轴上显示频数还是百分数由你选择 —— 使用哪种标度完全取决于你想凸显的主要事实。

折线图常用于显示随时间变化的数值。时间总是用横轴表示，频数用纵轴表示。通过在横轴上选择时间值，可以读出任何时间段内的频数，还能读出该时间点的相应频数。



折线图应只用于展现数值型数据，不应用于类别数据。原因是，对类别数据进行比较是有意义的，但为其绘制趋势线却没有意义。只有在基于某些数值型单位（比如时间）对类别进行比较时才使用折线图，这时，每一类别都用一条独立的线表示。



要点

- **累积频数**即到某个特定数值为止的总频数，即频数的累计总和。
- 通过累积频数图，可基于累积频数找出每组数据的上限。
- 需要体现趋势时请使用折线图，例如基于时间的趋势。
- 可用折线图显示多批数据。每批数据各用一条线表示，请确保能清楚识别每一条线。
- 由于通过折线图很容易看出趋势形状，因此可用折线图进行基本的预测。只要延长趋势线即可进行预测，但要尽量保持基本形状。
- **不要使用折线图显示类别数据**——除非要显示每一个类别的趋势，例如基于时间的趋势。如果要显示每一个类别的趋势，要为每一个类别画一条线。

世上没有傻问题

问： 折线图和时间序列图是一回事吗？我想我以前听到过这个名字。

答： 时间序列图确实是一种折线图。时间序列图以时间区间为关注点，我们用过的一些实例就是这样的。但折线图不一定要关注时间。

问： 折线图有什么特别的变体吗？

答： 有。事实上，你已经遇到过一种。**累积频数图**就是一种折线图，所显示的是到某个特定值为止的总频数。

问： 折线图既能显示类别数据，又能显示数值型数据吗？

答： 折线图显示类别数据的情况只有一种：只显示每一类别的趋势，且每条线代表一个类别。

折线图不应该用于这种情况：基于类别绘制线条。

问： 这么说在显示总体趋势时，折线图效果更好；在对数值或类别进行比较时，条形图效果更好？

答： 正确。使用哪种图形归根结底在于你要传递的信息，以及你要提炼的主要事实。

问： 既然我已经知道如何正确创建图形，我能用绘图软件完成这项繁重的工作吗？

答： 完全可以！绘图软件能为你节省大量时间，减少繁重工作，而且结果非常出色。

但要记住：软件无法代替你思考。你仍然需要决定哪种图能最好地体现你的主要事实，还必须检查软件所生成的结果是否正是你盼望得到的。

芒芒公司征服游戏市场！

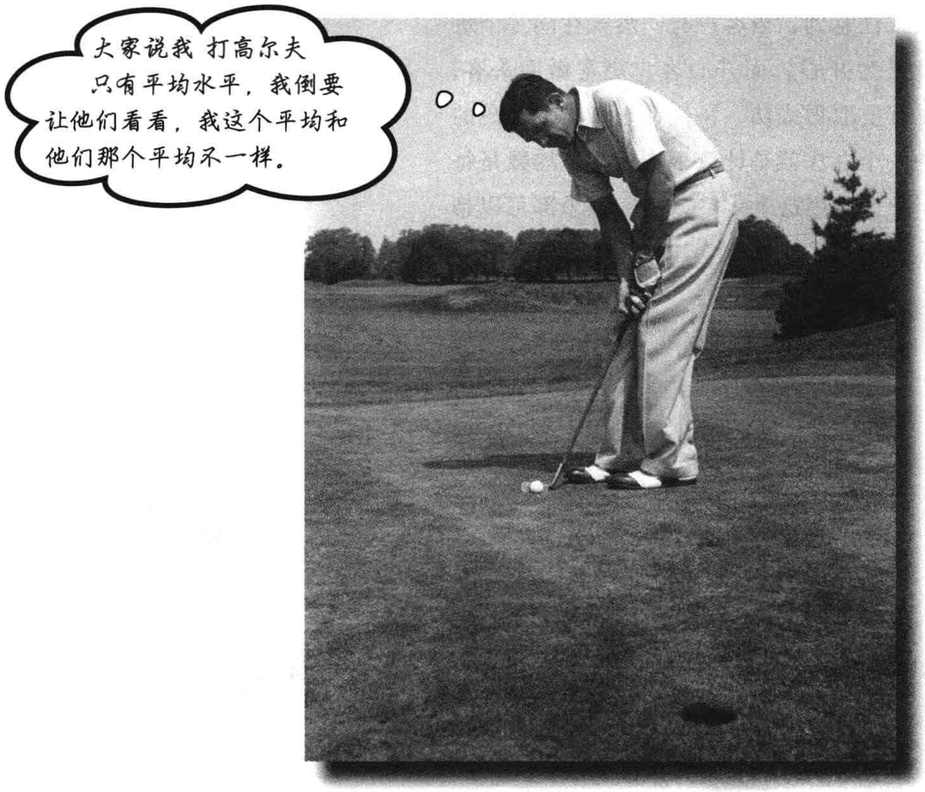
在你的帮助下，芒芒公司有了杀手锏，主题报告极为成功，这都是你的功劳。芒芒游戏名声大噪，赞助、广告纷至沓来。你唯一要做的就是想一想拿着大把的奖金干点什么，玩点什么。

统计学让你受益、对事情知根知底。你已经初尝甜头，接着读下去吧，我们将让你看到统计学能完成更多工作，你将真正开始让统计学发光发热。



2 集中趋势的量度

中庸之道



有时候，把握问题核心才是当务之急。

从一大堆数字中看出模式和趋势可能颇为不易，而求出**平均数**往往是把握全局的第一步。有了平均数就能迅速找出数据中最具代表性的数值，得出重要结论。在本章中，我们将介绍几种方法，帮助你计算最重要的统计量 —— 均值、中位数、众数。你将开始学习如何有效地**汇总数据**，尽可能得出简练、有用的结果。

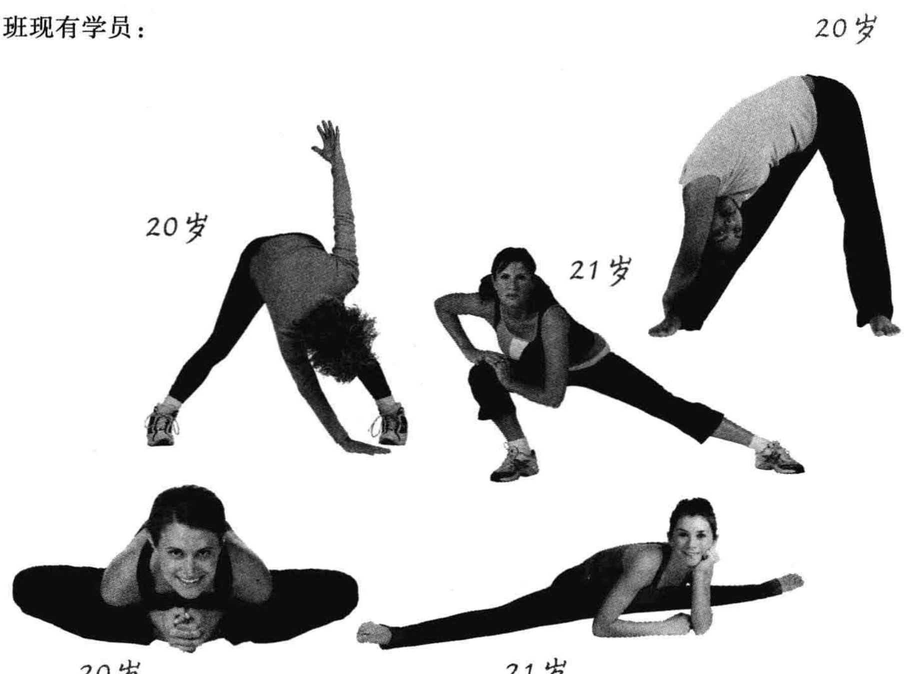
欢迎来到健身俱乐部

统计邦健身俱乐部深感自豪，因为他们有一项本事——能为每一位客户提供完美无缺的健身课程。无论你要学游泳、练武术，还是要打造型体，他们总有合适的课程等着你。

健身俱乐部
统计邦顶级养生馆

健身俱乐部的员工注意到，当客户与同龄人在同一个班上练习时，表现最为开心，而开心客户更常做回头客。看来，健身俱乐部要取得成功，秘诀在于算出每个班的典型年龄，其中一个办法就是**计算平均数**。平均数是每个班级的代表年龄，利用这个年龄，健身俱乐部可以帮助客户选择合适的班级。

下面是力量集训班现有学员：

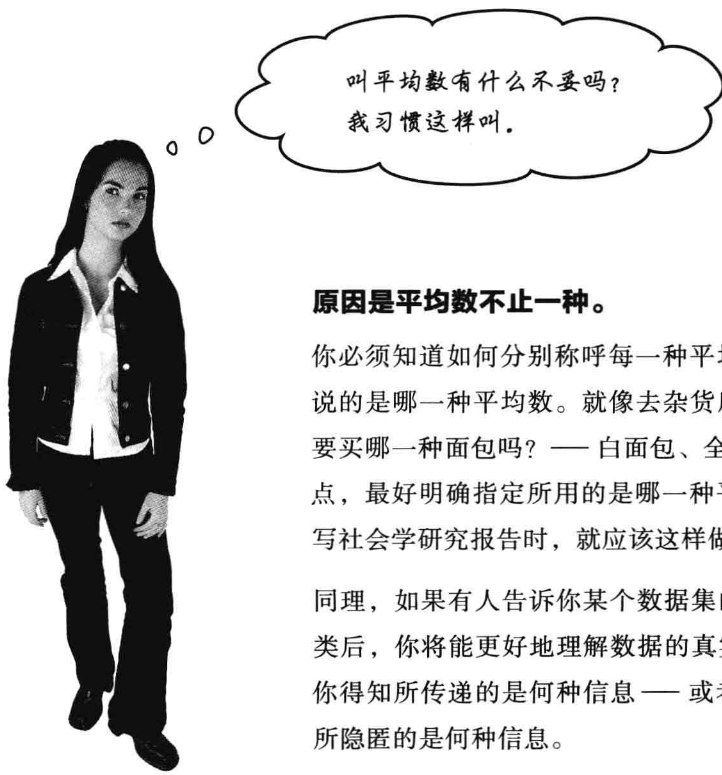


我们如何计算力量集训班的平均年龄？

均值：平均数的一般量度

可能以前有人让你算过平均数。计算大量数据的平均数的一个方法是：将所有数字加起来，然后除以数字个数。

在统计学中，这样算出来的值叫做**均值**。



原因是平均数不止一种。

你必须知道如何分别称呼每一种平均数，才能方便地告诉别人你所说的是哪一种平均数。就像去杂货店买面包，你不也得告诉售货员要买哪一种面包吗？—— 白面包、全麦面包或其他面包。考虑到这一点，最好明确指定所用的是哪一种平均数计算方法，例如，当你撰写社会学研究报告时，就应该这样做。

同理，如果有人告诉你某个数据集的平均数，当知道该平均数的种类后，你将能更好地理解数据的真实情况。这能给你重要线索，让你得知所传递的是何种信息 —— 或者，在某种情况下，会让你得知所隐匿的是何种信息。

我们先讲均值，随后在本章后面部分介绍其他类型的平均数。



均值数学

如果你想真正成为统计高手，就需要把一些常用统计符号用顺手。一开始可能会感觉有点儿生疏，但很快就会习惯的。

字母与数字

几乎每一种统计算法都涉及一批批数字的加法计算。例如，如果我们想求出力量集训班的年龄均值，首先就要把班上全体学员的年龄加起来。

统计师的问题是如何用通用方法表示这种算法。我们不一定事先知道有多少数字要处理，也不一定知道都有哪些数字。例如，我们目前知道力量集训班有多少人，知道他们的年龄，可要是有其他人加入，结果会怎么样？只有用通用方法表示以上算法，才有办法在班级情况发生变动时，不用重新推导，就能写出算法。

统计师是这样解决以上问题的：用字母表示数字。例如，他们可能会用字母**x**表示力量集训班中的学员年龄，如下所示：

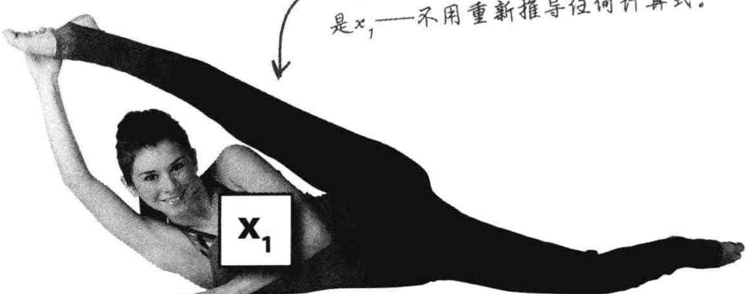
班级学员特定年龄

19 20 20 20 21 → **班级学员通用年龄** x_1 x_2 x_3 x_4 x_5

每个x表示班级中的一个人的年龄，有点儿像用特定数字x对每个人做标记一样。

每个x表示班级中的一个人。

我们用x_1表示这个女孩的年龄。她今年19岁，可是，就算她到了20岁，我们还是认为她的年龄就是x_1——不用重新推导任何计算式。



既然我们已经有了表示年龄的通用方法，就能用x进行各种计算。可以以下列方式表示班级中的5个人年龄的总和：

$$\text{Sum} = x_1 + x_2 + x_3 + x_4 + x_5$$

可要是我们不知道有多少数字需要求和该怎么办？例如，要是我们不知道班级中有多少人该怎么办？

处理未知条件

统计师用字母表示未知数字。可如果我们不知道有多少数字需要求和该怎么办？没问题——我们只要把这些数字的数目叫做*n*就可以了。例如，如果我们不知道力量集训班中有多少人，我们就说有*n*个人，然后将年龄和写为：

$$\text{Sum} = x_1 + x_2 + x_3 + x_4 + x_5 + \ldots + x_n$$

"…"是"以此类推"的简写，"以此类推"表示按序不断地增加x。

在本例中，x_n表示班上第n个人的年龄。如果班上有18个人，则这个数是x_{18}，即第18个人的年龄。



我们可以用另一种简捷表示法。

$x_1 + x_2 + x_3 + x_4 + \ldots + x_n$这种写法有点儿像在说"年龄1加年龄2，再加年龄3，然后加年龄4，依次类推，直到加到年龄n。"在日常交流中，我们不太可能这么说，而更可能说"把所有年龄加起来"，这样更直接、更简单、切中要点。

与此相似，在数学中，我们可以用Σ符号表示这个意思，Σ为希腊字母，读作"西格玛"。我们可以用Σx（读作：西格玛x）简捷地表示"将所有的x加起来"。

现在都加起来了……

$$x_1 + x_2 + x_3 + x_4 + x_5 + \ldots + x_n = \sum x$$

看到了吧，多直接、多简单啊！这就是"把所有数值加起来"的数学表示方法，不用明确说出每个数值。

讲过这些方便简单的数学表示法之后，让我们看看怎么用这种数学表示法计算均值。

再说均值

我们可以用数学符号表示均值。

为了求出一批数字的均值，我们会将这些数字加起来，然后除以这些数字的个数。我们已经讲过如何记总和，还讲过统计师如何用n来表示一批数字的总和。

把以上记法合并起来，均值就可以记为：

$$\frac{\sum x}{n}$$

把所有的数字加起来…

然后除以数字个数。

也就是说，这就是“将所有的数字加起来，然后除以数字个数”的简捷数学表示法。

均值的专用符号

均值是应用最广泛的统计量之一。由于使用如此频繁，统计师们专门给了它一个符号：μ。这是一个希腊字母（读作“缪”）。记住，这只是表示均值的一种简捷方法。

均值是应用最广泛的统计量之一，可用符号μ表示。

我是均值。有些人说我是平均数，但实质上，我叫均值。

$$\mu = \frac{\sum x}{n}$$

动动笔

试着算一下力量集训班的年龄均值？下面是学员们的年龄。

每种年龄的人数 →

年龄	19	20	21
频数	1	3	1

5分钟推理



案件：含含糊糊的平均数

本地一家公司的员工由于感到自己拿到的薪水不公道，出现了不满情绪。大部分员工周薪为500美元，少数经理高一些，而首席执行官每周搞回家49,000美元。

“这公司的平均薪水是每周2,500美元，而我们只有500。”工人们说，“这不公平，我们要加薪。”

一位经理耳闻了这个情况，也和他们一起要求加薪。“这公司的平均薪水是每周1万美元，而我只有4,000。我要加薪。”

首席执行官看着他们，说道：“你们都错了，平均薪水就是500美元一周，我没亏待谁，快回去干活吧。”

平均薪水是怎么回事？你认为谁是对的？

动动笔解答

试着算一下力量集训班的年龄均值？下面是学员们的年龄。

年龄	19	20	21
频数	1	3	1

为了求出μ，我们需要把所有人的年龄加起来，然后除以人数。即：

$$\mu = \frac{19 + 20 + 20 + 20 + 21}{5}$$

$$= \frac{100}{5}$$

$$= 20$$

记住，有3个人的年龄为20岁。

年龄均值为20。

处理频数

在计算一批数据的均值时，你常常会发现有些数字是重复的。只要看看力量集训班的年龄就知道，实际上有3个人的年龄是20岁。

有一点确实很重要：在计算均值的时候，要把每个数的**频数**考虑进去。为了确保自己不忽略这一点，我们可以把它写入公式。

如果用**f**代表频数，就可以重新将均值表示如下：

$$\mu = \frac{\sum fx}{\sum f}$$

每个数字乘以其频数，然后将全部乘积相加。

频数和

这是表示均值的另一种方法，但这次明确指出了频数。用这个方法计算力量集训班的数据，得出：

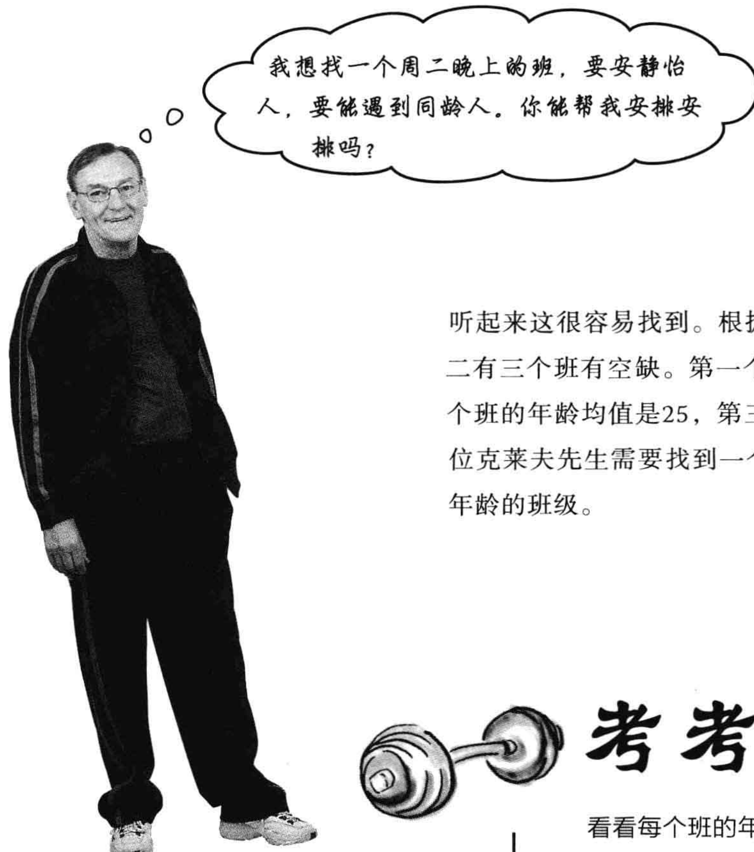
$$\mu = \frac{1 \times 19 + 3 \times 20 + 1 \times 21}{5}$$

$$= 20$$

计算方法相同，但写法略有区别。

再说健身俱乐部

又一位顾客满怀希望地前来寻找完美无缺的健身班。你能帮他找一个吗？



米特·克莱夫先生，年近花甲，想找一个由中年人组成的健身班。

听起来这很容易找到。根据宣传手册，健身俱乐部周二有三个班有空缺。第一个班的年龄均值是17，第二个班的年龄均值是25，第三个班的年龄均值是38。这位克莱夫先生需要找到一个学员平均年龄贴近他本人年龄的班级。

考考你

看看每个班的年龄均值。克莱夫应加入哪个班？

人人都在练功夫

克莱夫去了年龄均值为38岁的班。他盼望这是一个程度一般的班级，他可以在这里进行一些不太剧烈的练习，遇到一些年龄相仿的朋友。遗憾的是……



哪里出错了？

克莱夫曾经盼着加入的班级原来主要由十几岁学员组成。你觉得为什么会出现这种情况呢？

我们需要查看数据，探明究竟。让我们看看，草绘一个数据图，看是否有助于找出问题所在。

重要统计量

均值

$$\mu = \frac{\sum x}{n}$$

$$\mu = \frac{\sum fx}{\sum f}$$



练习

绘制功夫班和力量集训班的直方图（若要复习直方图，请参考第一章）。直方图的分布形状比较下来结果如何？克莱夫为什么会被分到错误的班级？

力量集训班学员年龄

年龄（岁）	19	20	21
频数	1	3	1

功夫班学员年龄

年龄（岁）	19	20	21	145	147
频数	3	6	3	1	1

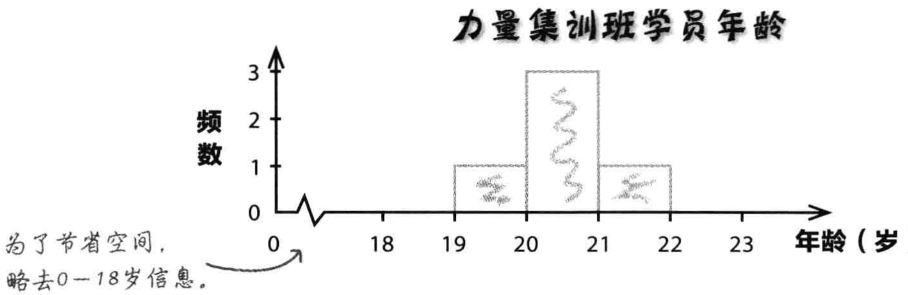
绘制功夫班和力量集训班的直方图（若要复习直方图，请参考第一章）。直方图的分布形状比较下来结果如何？克莱夫为什么会被分到错误的班级？

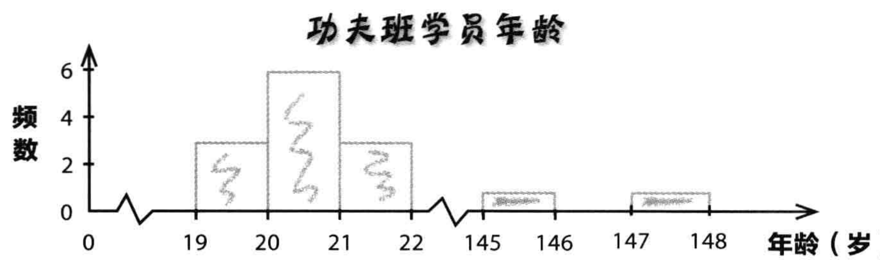
力量集训班学员年龄

年龄（岁）	19	20	21
频数	1	3	1

功夫班学员年龄

年龄（岁）	19	20	21	145	147
频数	3	6	3	1	1





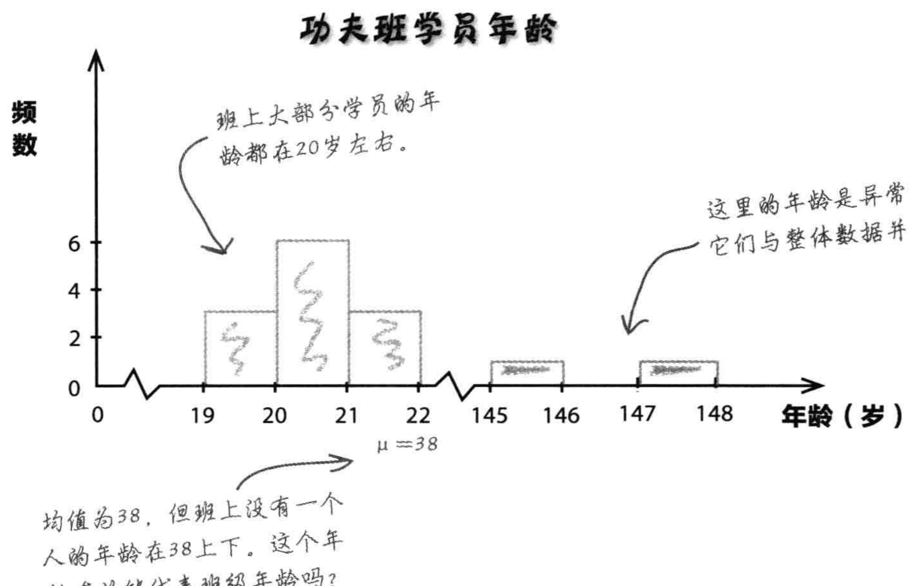
动动笔

你认为均值会是一批数据中的最大值吗？在什么情况下会是这样？

我们的数据中存在异常值

看出力量集训班和功夫班的图形形状有何差别了吗？力量集训班的年龄形成了光滑、对称的形状，很容易看出班上学员的典型年龄。

功夫班的图形形状则不这么直截了当。大部分年龄都在20岁左右，但有两位祖师爷的年龄远远超过20岁。像这样的极值被称为**异常值**。



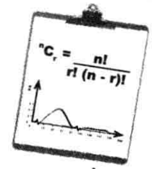
动动脑

如果这个班上不包括几位祖师爷，均值会是多少？将该均值与实际均值进行比较。你会因此得知异常值有何影响？

真凶是~~他~~异常值

观察功夫班的数据和图形，很容易看出班上学员的年龄在20岁左右。事实上，如果班上没有那几位祖师爷，20岁就是均值。

但我们不能简单地忽略那几位祖师爷：他们仍然是班上的一分子。遗憾的是，这几位明显高于“典型”年龄成员的存在扭曲了均值，使均值抬高了。

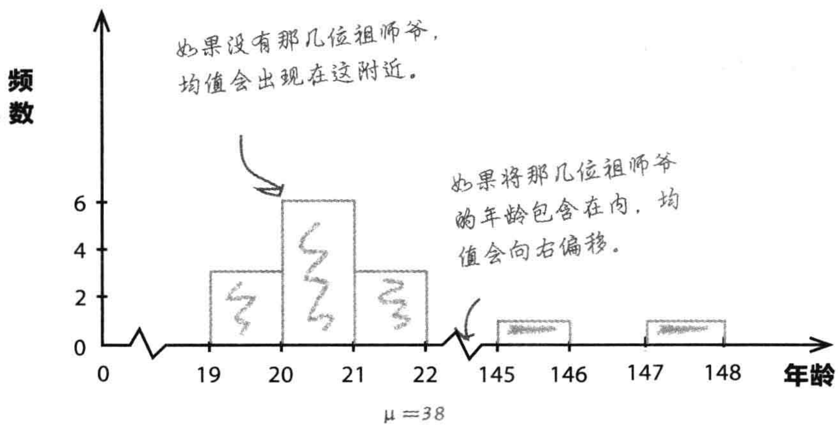


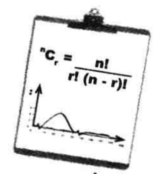
重要统计量

异常值

与其他数据格格不入的极高或极低的数值

功夫班学员年龄





重要统计量

偏斜数据

当异常值将数据向左或向右“拉”时即产生偏斜数据

你能看出异常值如何拉高均值吗？这就是异常值对数据的影响。一旦发生这种情况，我们就说数据**偏斜**了。

功夫班的数据向右偏斜，原因是，如果按照升序排列所有数据，异常值位于右边。

让我们仔细看看。



动动笔解答

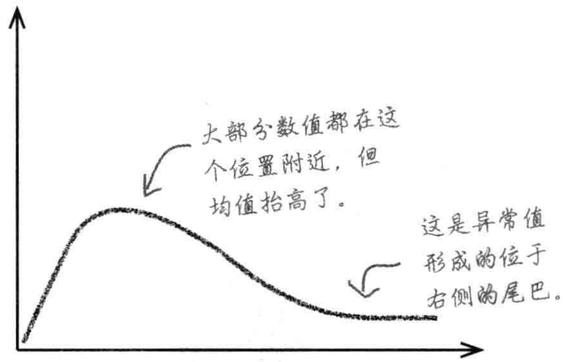
你认为均值会是一批数据中的最大值吗？在什么情况下会是这样？

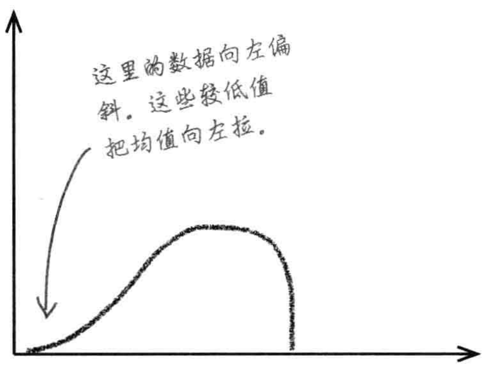
是的，会有这种情况。如果一批数据中的所有数据都相同，则均值会是最大值。

偏斜数据细细看

向右偏斜

向右偏斜的数据有一条“尾巴”，这条尾巴由偏大异常值形成，向右逐渐变弱。拿一张右偏斜图形看看，就能看到这样的尾巴。功夫班中的偏大异常值扭曲了均值，将均值拉高了——即拉向了右边。



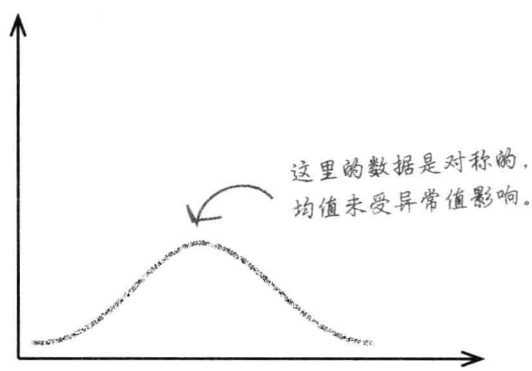


向左偏斜

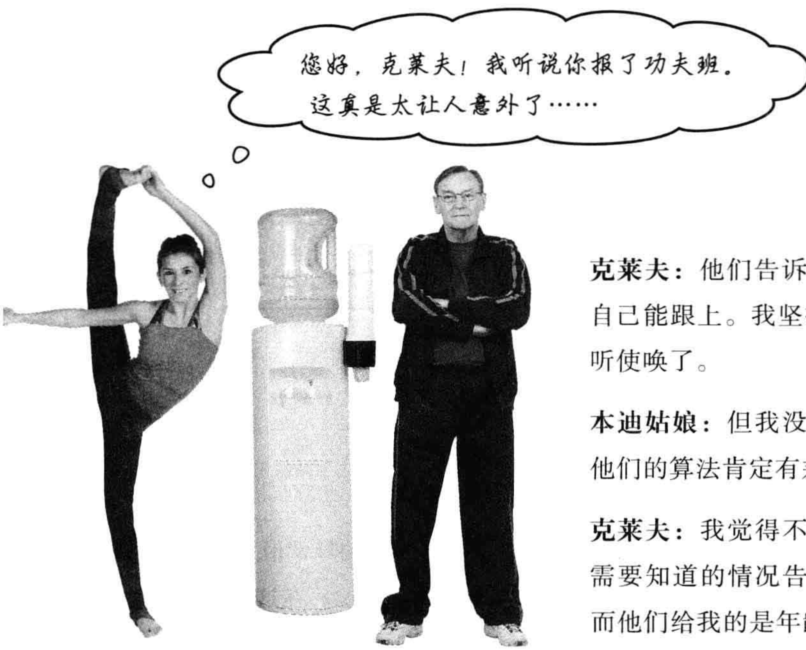
这张图上的数据向左偏斜。看到左侧的异常值尾巴了吗？这次的异常值位于低端，把均值向左拉。在这种情况下，均值小于大部分值。

对称数据

在理想情况下，你会希望数据呈对称形态。如果数据对称，则均值位于中央。不会有任何异常值将均值拉向任何一侧，中央位置两侧的数据形状大致相同。



饮水机边的对话



克莱夫：他们告诉我这个班的平均年龄是38岁，所以我觉得自己能跟上。我坚持了5分钟就不得不坐下，要不我的腿就不听使唤了。

本迪姑娘：但我没看到这个班上有任何人是这个年龄，所以他们的算法肯定有差错。他们为什么会那样跟你说呢？

克莱夫：我觉得不是他们的算法有错：他们只是没把我真正需要知道的情况告诉我。我问他们班上的典型年龄是多少，而他们给我的是年龄均值，38。

本迪姑娘：那并不是真正的典型值，对吗？我是说，仅看班上那些人的话，我会认为较年轻的年龄更具代表性。

克莱夫：要是他们把几位祖师爷从算法中剔除掉，我就会知道不该去这个班。原因就在这儿，我确信无疑。他们把整个算法都扭曲了。

本迪姑娘：好吧，如果几位祖师爷引起了这么大的问题，他们为什么不忽略这几位祖师爷呢？也许这样能得出更有代表性的班级年龄……

寻找中位数

当偏斜数据和异常值使均值产生误导时，我们就需要用其他方式表示典型值。我们可以取中间值，这种做法切实可靠。中间值是另一种平均数，我们称其为**中位数**。

为了求出功夫班的中位数：比如某个功夫班按升序排列所有年龄，取出中间值，如下所示：

19　19　20　20　20　21　21　100　102

这是中间的数字，即中位数，20。

如果把功夫班上的所有年龄按升序排列起来，数值20正好在当中。因此，功夫班的中位数为20。

要是班上学员数目为偶数该怎么办呢？

19　20　20　20　21　21　100　120

如果班上学员数目为偶数，则中间数并非仅有一个。

如果一批数字的数目是偶数，则只要取两个中间数的均值即可（将两个中间数加起来，再除以2），结果就是中位数。在上例中，中位数是20.5。

中位数永远处于中间，它是个中间值。

动动脑

我们已经看到，如果有9个数，则中位数是处于第5个位置的数；如果有8个数，则中位数是处于第4.5个位置（第4位和第5位中间）的数。要是有n个数呢？

求中位数三步法：

1. 按顺序排列数字：从最小值排列到最大值。

2. 如果有奇数个数值，则中位数为位于中间的数值。如果有n个数，则中间数的位置为(n+1)/2。

3. 如果有偶数个数值，则将两个中间数相加，然后除以2。中间位置的算法是：(n+1)/2。两个中间数分别位于这个中间位置的两侧。

世上没有傻问题

问： 如果确实想用均值，哪怕存在偏斜数据，还能用吗？

答： 可以用，而且大家经常这么做。不过，这时均值无法最恰当地体现典型值。你需要使用中位数。

问： 这是你的看法，但均值的主要意义的确是给出典型值，均值是个平均数。

答： 均值带来的巨大危险是：它会给出一个不存在于数据集中区的数值。以功夫班为例：如果你要加入这个班，并随机挑出一个人，很可能这个人是在20岁左右，因为班上大多数人的年龄都在20岁左右——只看均值无法形成这种印象，求出中位数会让你对数据有更准确的预期。

但即使是中位数，有时也会得出不存在于数据集中区的值，上一页的例子就是这样。这正是出现多种平均数的原因，有时候，为了正确地指出典型值，需要使用各种各样的方法。

问： 这么说中位数比均值更好？

答： 有时候中位数比均值更合适，但这并不是说它更好。大多数时候，你会需要使用均值，因为均值的优势通常远胜中位数，均值对于抽样数据来说更稳定。本书后文会继续阐述这一点。

问： 对于类别数据该怎么使用均值或中间值呢？对于一些实例，像第1章第9页中的数据，该怎么办？

答： 你只能求数值型数据的均值和中位数。不过别担心，还有一种平均数可以处理这种问题，我们随后会展开讲。

问： 我总是搞不清右偏斜数据和左偏斜数据。怎样才能记住哪是右偏斜，哪是左偏斜？

答： 偏斜数据有一条“异常值”尾巴。若要知道数据的偏斜方向，可看看尾巴的指向。例如，右偏斜数据的尾巴指向右方。



化身为数据

请假装成数据来玩这个游戏，说一说每个数据集的中位数是哪一个、数据是否偏斜、均值是大于还是小于中位数。

请说出理由。

数值	1	2	3	4	5	6	7	8
频数	4	6	4	4	3	2	1	1

数值	1	4	6	8	9	10	11	12
频数	1	1	2	3	4	4	5	5

化身为数据



请假装成数据来玩这个游戏，说一说每个数据集的中位数是哪一个、数据是否偏斜、均值是大于还是小于中位数。

请说出理由。

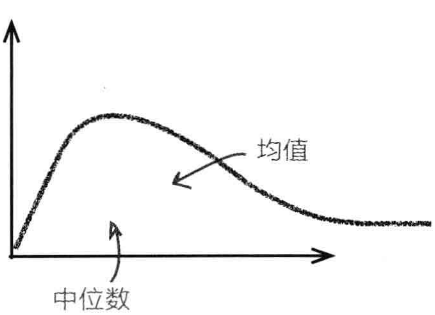
数值	1	2	3	4	5	6	7	8
频数	4	6	4	4	3	2	1	1

这里有25个数，如果把这些数排列起来，中位数正好在中间，即在第13个数的位置；中位数为3；数据向右偏斜；均值被拉高，因此，均值大于中位数。

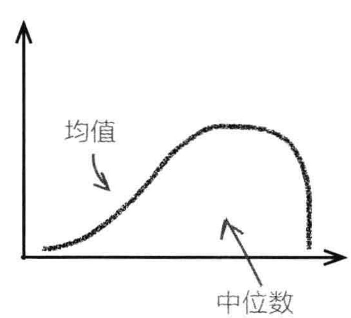
数值	1	4	6	8	9	10	11	12
频数	1	1	2	3	4	4	5	5

这里的中位数为10；数据向左偏斜；均值被拉向左边，因此，均值小于中位数。

如果数据向右偏斜，则均值位于中位数右侧（较大）。



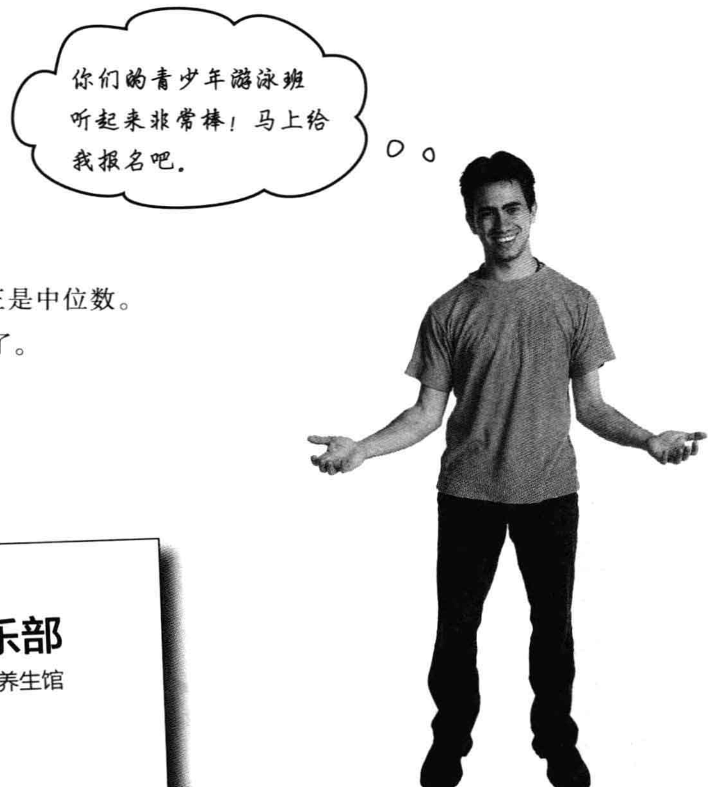
如果数据向左偏斜，则均值位于中位数左侧（较小）。



生意日益兴隆

你对平均数的研究的确得到了回报，越来越多的人前来健身俱乐部挑选健身班，员工们发现，为客户们挑选合适的班级变得容易多了。

这位十几岁的小青年正在找游泳班，他想在班上交一些年龄相仿的新朋友。



游泳班的年龄均值是17，巧的是，这正是中位数。
听上去，这个班对于他来说再合适不过了。



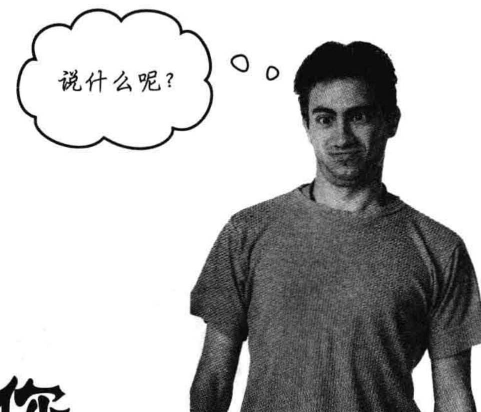
让我们看看故事的发展……

小鸭呱呱游泳班

小鸭呱呱游泳班每周在游泳池里碰头两次。在这里，家长们教他们的小宝宝学游泳，大家玩水嬉戏，乐不可支。

看看谁来上课了……





考考你

你认为这一次又是哪里出差错了呢？



掉落的频数磁贴

下面是参加小鸭呱呱游泳班的成员的年龄，但有一些写有频数的磁贴掉下来了。你的任务是将这些频数放回频数表中的正确位置。参加这个班的有9个孩子及其父母，均值和中位数都是17。

年龄	1	2	3	31	32	33
频数	3		2	2		



动动笔

弄清楚小鸭呱呱游泳班的频数后，画出直方图。你注意到什么了？

掉落的频数磁贴

下面是参加小鸭呱呱游泳班的成员的年龄，但有一些写有频数的磁贴掉下来了。你的任务是将这些频数放回频数表中的正确位置。参加这个班的有9个孩子及其父母，均值和中位数都是17。

年龄	1	2	3	31	32	33
频数	3	4	2	2	4	3

已知有9个孩子，因此孩子的频数加起来肯定是9。肯定有4个2岁的孩子。

均值为17。如果我们用a和b表示未知频数，则：

$$\frac{1\times3 + 2\times4 + 3\times2 + 31\times2 + 32a + 33b}{18} = 17$$

两边都乘以18。

$$3 + 8 + 6 + 62 + 32a + 33b = 17\times18 = 306$$

$$32a + 33b = 306 - (3 + 8 + 6 + 62) = 306 - 79$$

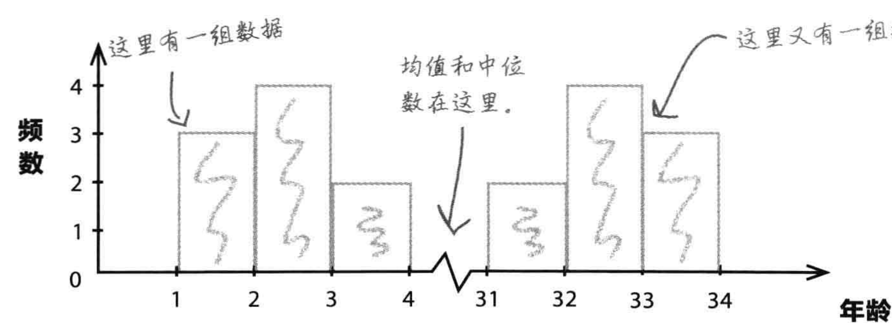
$$32a + 33b = 227$$

由于$32a+33b$是奇数，所以b肯定是3，a肯定是4。

动动笔解答

弄清楚小鸭呱呱游泳班的频数后，画出直方图。你注意到什么了？

小鸭呱呱班的学员年龄



看起来这不是一批数据，而是两批：一批是父母的，一批是孩子的。

均值和中位数出了什么问题？

让我们更细心地看看情况。

下面是参加小鸭呱呱游泳班的成员的年龄。

1 1 1 2 2 2 2 3 3 | 31 31 32 32 32 32 33 33 33

数字个数为偶数，因此中位数居于3和31当中。取这两个数的均值：(3+31)/2，得到17。

虽然班上没有一个人是17岁，但这个班级的年龄均值和中位数都是17！

可如果班上人数是偶数会怎么样呢？均值和中位数仍然具有误导性。请看：

1 1 1 2 2 2 2 2 3 (3) 31 31 32 32 32 32 33 33 33

如果班上再增加一个2岁的人，则中位数为3。那么成年人又怎么解释呢？

如果班上再增加一个2岁的孩子，如上所示，中位数仍然是3。这反映出孩子的年龄，但没有将成年人考虑在内。

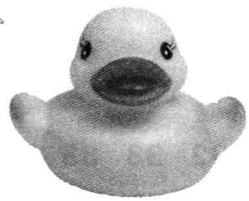
1 1 1 2 2 2 2 2 3 (31) 31 31 32 32 32 32 33 33 33

如果我们在班级中再增加一个31岁的成年人，中位数就会变为31。这一次，我们忽略了孩子！

如果再在班级中增加一个33岁的人，则中位数变为31。但这无法反映班上所有孩子的情况。看来，无论我们选择哪一个值作为平均年龄，总会出现误导。

我们该怎么处理这样的数据呢？

动动笔



现在请认认真真地考虑如何以最佳方式表示小鸭呱呱游泳班的代表年龄。下面是数据提示：

年龄	1	2	3	31	32	33
频数	3	4	2	2	4	3

1. 为什么你认为均值和中位数都不适用于这些数据？为什么均值和中位数具有误导性？

2. 如果必须挑选一个年龄来代表这个班级的年龄，这个年龄是多少？为什么？

3. 要是能挑选**两个**年龄呢？你会挑选哪两个年龄？为什么？



均值访谈

本周话题：
平均数的各种形式

Head First： 你好，平均数，很高兴邀请你来参加节目……

均值： 拜托，叫我均值。

Head First： 均值？可我想你是平均数。我们搞错来宾名单了吗？

均值： 完全没有。要知道，统计邦中的平均数不止一种，我是其中一种，叫作均值。

Head First： 平均数不止一种？听起来有点儿复杂。

均值： 其实不复杂，用习惯就好了。你看，我们都表示一批数字的典型值，但对于这个典型值是多少，我们各有各的看法。

Head First： 那么你们当中谁是真正的平均数呢？我说的是把所有数字加起来，然后除以数字个数所得到的那个？

均值： 是我。不过请别叫我“真正”的平均数，其他兄弟可能会恼火。真实情况是，大多数刚来统计邦的人都把我当作“平均数先生”，我的计算方法和学生们在基本算术中首次接触平均数时用的计算方法相同。只有在统计邦，我才叫做均值，以便和其他类型的平均数区分开来。

Head First： 那么你有其他名字吗？

均值： 说起来我确实有一个符号：μ。所有的摇滚明星都有别名，呃，一部分明星有别名，好歹我也有。这是个希腊名字，这让我颇具异国情调。

Head First： 那么为什么还需要别的平均数呢？

均值： 我讨厌承认这一点：我有缺点。当我处理存在异常值的数据时，就会变得没头没脑。没有异常值的时候，我表现很好，但只要看到异常值，我就会失魂落魄地跟着这些异常值走。这会带来不少问题。有时候我会远远偏离大部分数值所在的位置。这时就该请中位数出面了。

Head First： 中位数？

均值： 碰到异常值的时候，他真是太冷静了。无论你砸给他什么数据，他总是能端端正正地站在中间。当然了，中位数有他不好的一面：他无法计算。你只能指出他应该出现在哪个位置。随着计算深入，他的作用会有所逊色。

Head First： 你们二位有数值相等的时候吗？

均值： 如果数值是对称的，我们就会数值相同，否则我们往往不相同。一般规律是，如果存在异常值，那么我往往朝着异常值移动，而中位数则停在原来的地方不动。

Head First： 时间快到了，最后再问一个问题：会不会有这样的情况，用你和用中位数表示典型值都会出现问题？

均值： 恐怕有这种情况。有时候我们需要稍微借助另一种类型的平均数。他露面不是太多，但认识认识他很有用。别急，我将让你看看他都忙些什么。

Head First： 好极了！

动动笔解答

现在请认认真真地考虑如何以最佳方式表示小鸭呱呱游泳班的代表年龄。下面是数据提示：

年龄	1	2	3	31	32	33
频数	3	4	2	2	4	3

1. 为什么你认为均值和中位数都不适用于这些数据？为什么均值和中位数具有误导性？

对于以上数据，均值和中位数都具有误导性，因为两者都没有全面表示出班级中的成员的典型年龄。均值说明有一些十几岁的青少年参加了游泳班，实际上一个也没有，中位数也有同样的问题，但如果有别的人加入班级，中位数会大幅度波动。

2. 如果必须挑选一个年龄来代表这个班级的年龄，这个年龄是多少？为什么？

的确不太可能挑出一个完全代表班级年龄的年龄。这个班级实际上是由两批年龄组成的：一批是孩子的年龄，一批是家长的年龄。确实无法用一个数字同时代表两批年龄。

3. 要是能挑选两个年龄呢？你会挑选哪两个年龄？为什么？

由于这些数据看上去包括两批数据，挑选两个年龄来代表班级年龄是有意义的，一个年龄代表孩子们的年龄，一个年龄代表家长们的年龄。我们会选择2和32，因为这两个年龄组的成员最多。

认识众数

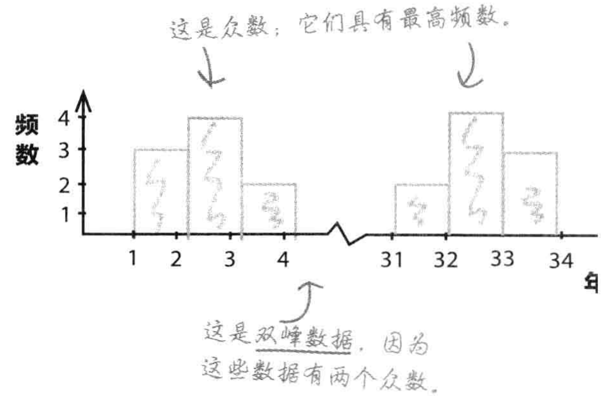
除了均值和中位数，还有第三种平均数，称为**众数**。众数是一批数字中最常见的数值，即频数最大的数值。与均值和中位数不同，众数**必须**是数据集中的一个数值，而且是最频繁出现的数值。

年龄	1	2	3	31	32	33
频数	3	4	2	2	4	3

这两个数最常出现，因此二者都是众数。

有时候，数据的众数可以不止一个。如果有一个以上的数值具有最大频数，则每一个这样的数值都是众数。如果数据看上去体现了多种趋势或多批数据，那么我们就为每一批数据给出一个众数。如果一批数据有两个众数，则我们说这种数据是**双峰**数据。

小鸭呱呱班的学员年龄



这正是我们在小鸭呱呱游泳班碰到的情况。我们的确观察到了两批数据，一批是家长的，一批是孩子的，因此不存在某一个能完全代表整个班级的年龄。相反，我们可以看出每一批年龄的众数。在小鸭呱呱游泳班上，年龄2和年龄32出现的频率最高，因此这两个年龄都是众数。从图上看，众数就是具有最高频数的年龄。

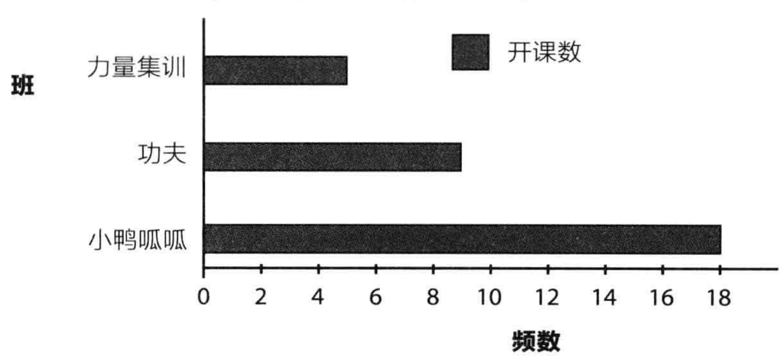
众数甚至能用于类别数据

众数不仅能用于数值型数据，还能用于类别数据。事实上，众数是**唯**一能用于类别数据的平均数。在处理类别数据时，众数是最常出现的平均数类型。

你还可以用众数指定具有最高频数的数值**组**。具有最高频数的组被称为**众数组**。



众数甚至能用于类别数据



求众数三步法：

1. 把数据中的不同类别或数值全部找出来。

2. 写出每个数值或类别的频数。

3. 挑出具有最高频数的一个或几个数值，得出众数。

动动笔

求出以下几批数据的众数。

数值	1	2	3	4	5	6	7	8
频数	4	6	4	4	3	2	1	1

类别	蓝	红	绿	粉	黄
频数	4	5	8	1	3

数值	1	2	3	4	5
频数	2	3	3	3	3

你认为众数在什么情况下最有用？

众数在什么情况下最无用？

恭喜！

你在健身俱乐部的辛勤工作正迎来巨大的成功，要求报班的人热情高涨。



动动笔解答

求出以下几批数据的众数。

数值	1	2	3	4	5	6	7	8
频数	4	6	4	4	3	2	1	1

这里的众数是2，因为2具有最高频数。

类别	蓝	红	绿	粉	黄
频数	4	5	8	1	3

这里的众数为“绿”。

数值	1	2	3	4	5
频数	2	3	3	3	3

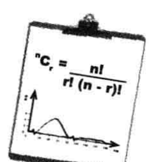
这一批数据有好几个众数：2，3，4，5。

你认为众数在什么情况下最有用？

当众数的数目较少时，或者，当数据为类别数据，而不是数值型数据时。均值和中位数都不能用于类别数据。

众数在什么情况下最无用？

当众数很多时。



重要统计量

众数

众数必须存在于数据集中。
众数是唯一能用于类别数据的平均数。



填写下表，针对我们在本章遇到过的平均数，写出算法，然后指出在哪种情况下会使用哪种平均数。请尽最大努力填写，不要回头翻阅本章的内容。

平均数	**计算方法**	**何时使用**
均值(μ)		在数据非常对称，且仅显示出一种趋势时使用。
中位数		
众数		



填写下表，针对我们在本章遇到过的平均数，写出算法，然后指出在哪种情况下会使用哪种平均数。请尽最大努力填写，不要回头翻阅本章的内容。

平均数	计算方法	何时使用
均值(μ)	以下任一算法均可 $\frac{\sum x}{n}$ ← x为每一个数值。← n为数值数目。 或 $\frac{\sum fx}{\sum f}$ ← f是每个x的频数。	在数据非常对称，且仅显示出一种趋势时使用。
中位数	将所有数据按照升序顺序进行排列。如果有奇数个数值，则中位数为中间的数值；如果有偶数个数值，则中位数为两个中间的数值相加再除以2得到的结果。	在数据由于异常值而发生偏斜时使用。
众数	选出具有最大频数的一个或几个数值。如果数据可分为两组，则为每组找出一个众数。	在遇到类别数据时使用。 当数据可以分为两个或更多组时使用。 众数是唯一能用于类别数据的平均数类型。

动动笔

星巴仕咖啡连锁店慷慨大方的首席执行官想给全体员工加薪。他不太确定，是直接给每个人加2,000美元呢，还是按10%的比例加。薪水均值为50,000美元，中位数为20,000，众数为10,000。

a) 如果星巴仕每位职员都加薪2,000美元，均值、中位数和众数都会发生哪些变化？

b) 如果星巴仕每位职员都加薪10%，均值、中位数和众数都会发生哪些变化？

c) 如果你的薪水为均值，你希望采用哪种加薪方式？如果你的薪水等于众数呢？

动动笔解答

星巴仕咖啡连锁店慷慨大方的首席执行官想给全体员工加薪。他不太确定，是直接给每个人加2,000美元呢，还是按10%的比例加。薪水均值为50,000美元，中位数为20,000，众数为10,000。

a) 如果星巴仕每位职员都加薪2,000美元，均值、中位数和众数都会发生哪些变化？

均值：如果x代表原来的薪水，n代表员工数目：

$$\mu = \frac{\sum(x + 2000)}{n}$$

$$= \frac{\sum x}{n} + \frac{\sum 2000}{n}$$

$$= 50,000 + \frac{2000n}{n}$$

$$= \$52,000$$

原来的均值

薪水为2,000的有n人次。

每个人的薪水都增长2,000美元会令均值、中位数和众数都增长2,000美元。

均值：每一份薪水都增加2,000美元，中间值（即中位数）也是如此。新的中位数是：

$20,000 + $2,000 = $22,000.

众数：最常见的薪水（或者叫做众数）为10,000美元，当增加2,000美元后，众数变为：

$10,000 + $2000 = $12,000.

b) 如果星巴仕每位职员都加薪10%，均值、中位数和众数都会发生哪些变化？

这一次，所有的薪水都乘以1.1（即100% + 10%）。

均值：

$$\mu = \frac{\sum(1.1x)}{n}$$

$$= \frac{\sum 1.1\,x}{n}$$

$$= 1.1 \times 50,000$$

$$= \$55,000$$

每个人加薪10%，则均值、中位数和众数也增加10%。

中位数：每一份薪水都乘以1.1，中间数（即中位数）也是如此。新的中位数为：

$20,000 × 1.1 = $22,000.

众数：最常见的薪水（或者叫做众数）为10000美元，众数乘以1.1后，变为：

$10,000 × 1.1 = $11,000.

c) 如果你的薪水为均值，你希望采用哪种加薪方式？如果你的薪水等于众数呢？

如果你拿的薪水是均值，则加薪10%的加薪幅度更大；如果你拿的薪水是众数，则直接加薪2,000美元的加薪幅度更大。

破案：含含糊糊的平均数

平均薪水是怎么回事？你认为谁是对的？



工人、经理和首席执行官各自用了不同的平均数。工人们用了中位数，这使得首席执行官的薪水造成的影响达到最低程度。

经理们用了均值。首席执行官的高薪令数据向右偏斜，均值因此显得虚高。

首席执行官用了众数。大部分工人的薪水为每周500美元，所以500美元就是薪水的众数。

那么，谁对谁错？从某种意义上说，他们都是对的，但我们不得不说，每一个人群都在使用最有利于自己意愿的平均数。记住，统计量能够提供信息，但也能造成误导。权衡再三，我们认为最适合用于本案例的平均数是中位数，因为数据中存在异常值。

3 分散性与变异性的量度

强大的“距”



世事可靠不可靠，我们该问谁？

平均数在寻找数据集典型值方面十分了得，**但平均数并不能说明一切。**平均数能让你知道数据中心所在，但若要给数据下结论，仅有均值、中位数和众数往往无法提供充足信息。在本章中，我们将开始分析**各种距和差**，让你的数据分析技术进入新境界。

招聘：队员一名

统计邦全明星篮球队是当地炙手可热的篮球队，是今年联赛的夺冠热门。只是，由于一场离奇的意外事故，他们有一位队员倒下了。他们需要一名新队员，越快越好。

新队员必须是全才，但教练真正需要的是一位靠得住的投篮手。只要球员取得他的信任，使他相信球员有能力投篮得分，他就会成为篮球队的一员。

教练整整一星期都在试用球员，他发现有三位球员可以考虑。问题是，他该选择哪一位？



我们需要比较球员得分

下面是三位球员的得分：

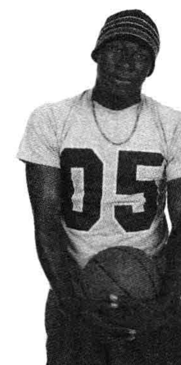


每场比赛的得分	7	8	9	10	11	12	13
频数	1	1	2	2	2	1	1

此处的频数告诉我们球员获得每种得分的比赛场数。这位球员有2场比赛得9分，有1场比赛得12分。



每场比赛的得分	7	9	10	11	13
频数	1	2	4	2	1



每场比赛的得分	3	6	7	10	11	13	30
频数	2	1	2	3	1	1	1

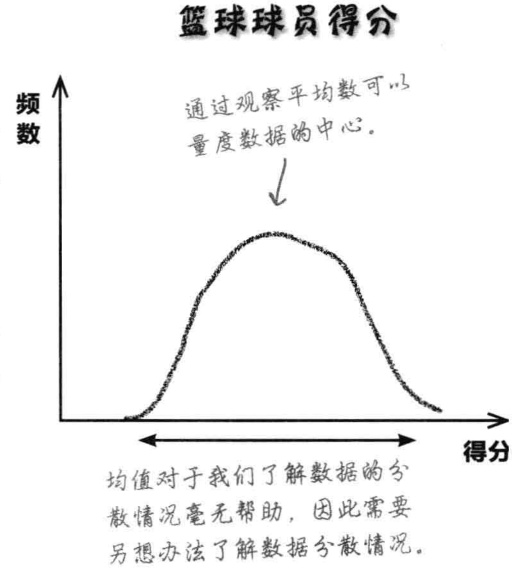
每位球员的得分均值、中位数和众数都是10分，但只要你注意一下所有得分就会发现，这几位球员是以不同的方式获得这些成绩的。球员们在稳定发挥方面存在差异，平均数无法量度这一差异。

我们需要通过某种方法对三人的得分进行分析，以便为球队挑选出最合适的人选。除了平均数，我们还需要用其他方法对数据进行比较——用哪一种方法呢？

动动脑

除了平均数以外，还有哪种信息会帮助教练作出决定？

使用全距区分数据集



前面讲过数据集平均数的计算方法，但平均数往往只给出部分信息。平均数让我们有办法确定一批数据的中心，却无法知道数据的变动情况。在前面的例子中，虽然每一位球员的平均得分相同，但显然各个数据集之间存在差异，我们需要通过某种方法量度这些差异。

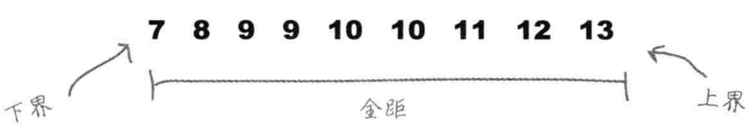
我们可以观察球员得分相对于平均数的分散情况，以此区分各个数据集。每位球员的得分分布情况各不相同，只要能够量度这些得分的分布情况，教练就能够做出更有依据的决策。

量度全距

通过计算全距（也叫极差），我们可以轻易获知数据分散情况。**全距**指出数据的扩展范围，有点儿像测量数据的宽度。全距的计算方法是：用数据集中的最大数减去数据集中的最小数。

最小值称为**下界**，最大值称为**上界**。

让我们看看其中一个球员的得分，再看看如何运用全距。下面是得分：



为了计算全距，我们用上界减下界。从数据中看出，最小值为7，因此这是下界；同样可以看出上界，即最大值13。用上界减下界，得到：

全距 = 上界 − 下界
= 13 − 7
= 6

所以该数据集的全距为6。

全距是量度数据分散程度的既简单又方便的方法，于是，我们有了另一种对数据集进行比较的方法。

重要统计量

全距

全距也叫极差，是用于量度数据集分散程度的一种方法。其算法为：

上界−下界

其中上界为最大值，下界为最小值。



练习

算出下列数据的均值、下界、上界、全距，画出图形。数值的分布方式相同吗？全距能否帮助我们描述这些差异？

得分	8	9	10	11	12
频数	1	2	3	2	1

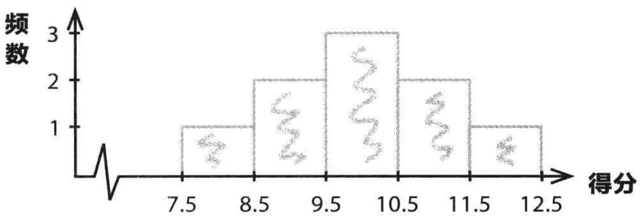
得分	8	9	10	11	12
频数	1	0	8	0	1



练习解答

算出下列数据的均值、下界、上界、全距，画出图形。数值的分布方式相同吗？全距能否帮助我们描述这些差异？

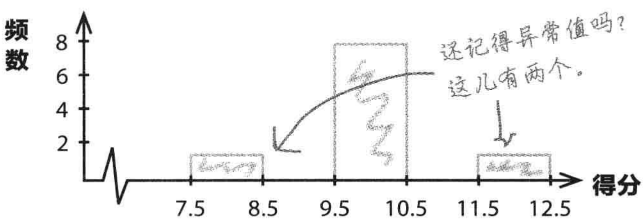
得分	8	9	10	11	12
频数	1	2	3	2	1



$\mu = 10$
下界 = 8
上界 = 12
全距 = 12 − 8
= 4

看，数据各不相同，这些计算结果却一样。

得分	8	9	10	11	12
频数	1	0	8	0	1



$\mu = 10$
下界 = 8
上界 = 12
全距 = 12 − 8
= 4



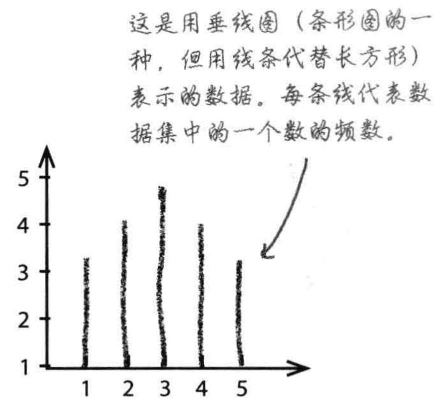
以上两个数据集的全距相同，但数值分布情况却有差别。我在想，全距是否确实包含有关数据分散情况的全部信息？

全距仅仅描述了数据的宽度，并没有描述数据在上、下界之间的分布形态。

以上两个数据集都具有相同的全距，但第二个数据集有异常值（即极大值和极小值）。看来，全距能量度数值的展开宽度，但很难得出数据的真实分布形态。

异常值带来的问题

全距是描述数据集分散程度的简便方法，但通常并非描述数据在该全距内的分布形态的最好方法。如果你的数据中包含异常值，那么，使用全距描述数据的分散情况会极具误导性，原因是全距很容易受异常值影响。让我们看看具体情况。



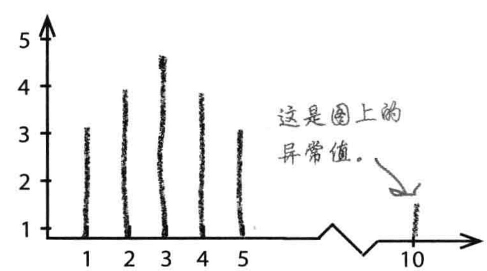
假想我们有以下一批数据：

下界1　　　　　　　　　　上界5

1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 5 5 5

这里的数字非常均匀地分布在上界和下界之间，并且无需担心任何异常值。这一批数据的全距为4。

可要是增加一个异常值，例如10，会发生什么变化呢？



下界还是1。　　　　　　　　但上界增长到10。

1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 5 5 5 10

下界保持不变，但上界增加至10，于是新全距为9。仅仅因为额外增加了一个数——一个异常值，全距就增长了5。

没有这个异常值的时候，以上两批数据是相等的，那么，我们对数值分布形态的描述为什么会出现这样大的差别呢？

动动脑

你能不能想个办法，我们按照这个办法构建一个距，使这个距受异常值影响不大？



全距是表述数值分布情况的一种极其简单方便的办法，但颇有一些局限性。

全距指出数据最大值和最小值之间的差距，但仅此而已 —— 全距只是对数据分布情况极其基本的描述。

全距的主要问题是：仅仅描述了数据的宽度。由于全距是通过数据极值计算得出的，因此不可能指出数据的真实形态以及数据是否包含异常值。构成相等全距的途径很多 —— 有时候这一点附加信息十分重要。



主要原因是全距非常简单。

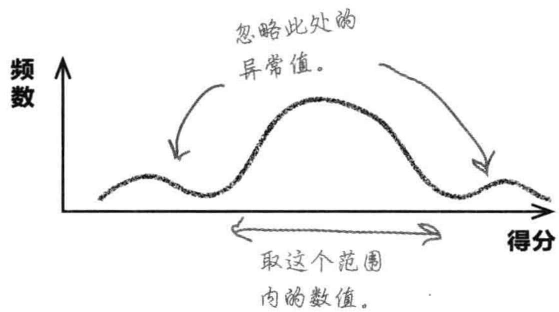
全距如此简单，大家都能理解 —— 即使很少接触统计学的人也不例外。例如，当你谈起年龄全距时，大家很容易就能理解你的意思。

不过，请小心，在全距极其简单的表象下却潜伏着危机。由于全距无法反映最大值和最小值之间的详细情形，使用时很容易让人对基础数据产生误会。

我们需要摆脱异常值

从全距的定义可以看出，全距的主要问题是包含异常值。只要数据中有异常值，即使只有一两个，全距中就会包含这些异常值。我们需要通过某种方法消除这些异常值的影响，这样才能最好地描述数据的分布形态。

有一个办法可以解决这个问题，即使用所谓的**迷你距**忽略异常值。我们不再量度整个数据集的全距，而是找出这个全距的一个部分——不包含异常值的部分。





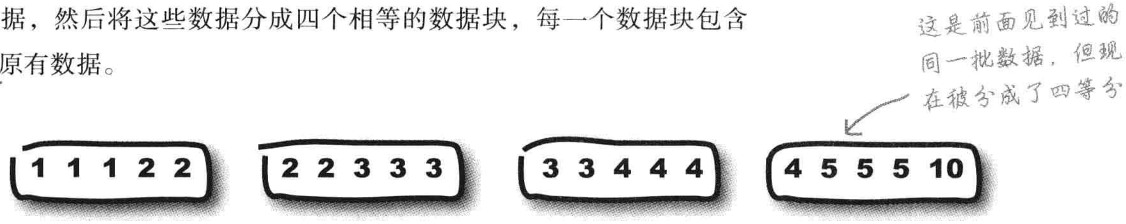
我们需要用一个统一的方法摆脱异常值。

如果随心所欲地忽略异常值，会产生这样一个问题：很难对几个数据集进行比较——谁知道是不是所有数据集都以完完全全相同的方式忽略了异常值？

我们需要确保这一点：对要进行比较的几个数据集统统使用相同的迷你距定义。如何办到呢？

四分位数出手相救

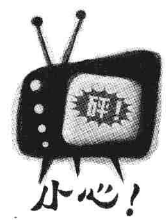
构建迷你距的一个办法是：仅使用数据中心周边的数值。为此，首先按升序排列数据，然后将这些数据分成四个相等的数据块，每一个数据块包含四分之一原有数据。



我们可以用介于两条外分割线之间的数值构建一个距。



如上，起到将整批数据一分为四作用的几个数值就是所谓的**四分位数**。求四分位数的方法有点儿类似求中位数，不同之处在于，需要求出将整批数据一分为四的几个数值，而不是求出将整批数据一分为二的一个数值。



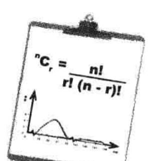
有一些教材在提到四分位数时，指的是每一份四分之一数据块中的所有数。

我们不是这样。我们用术语**四分位数**特指将整批数据一分为四的几个数值。

最小的四分位数（Q1）称为**下四分位数**或第一四分位数，最大的四分位数（Q3）称为上四分位数或第三四分位数。中间的四分位数（Q2）就是中位数，因为它将数据一分为二。每两个四分位数之间的距被称为**四分位距**（IQR）。

四分位距 = 上四分位数 – 下四分位数

四分位距为我们提供了一种用于量度数据分散程度的标准的、可重复使用的方法，这是另一种能对数据进行比较的方法。但异常值会怎么样呢？四分位距也能帮助我们处理异常值吗？让我们看一看。



重要统计量

四分位数

四分位数是这样一些数值：它们将数据一分为四。最小的四分位数称为下四分位数，最大的四分位数称为上四分位数。

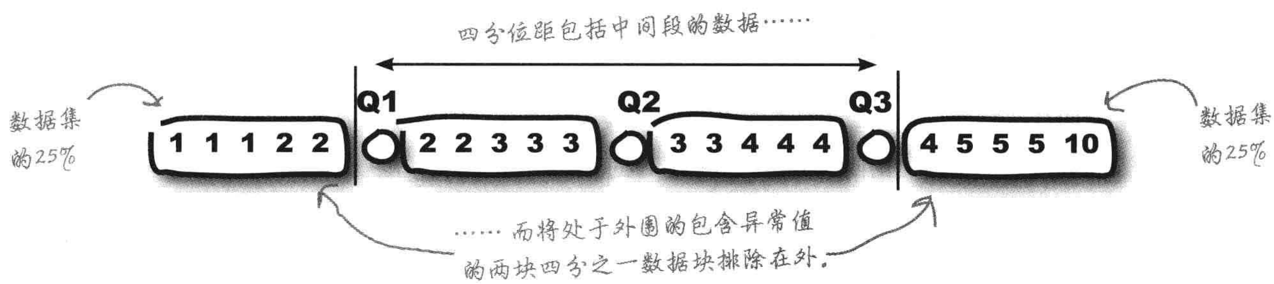
中间的四分位数即中位数。

四分位距剔除异常值

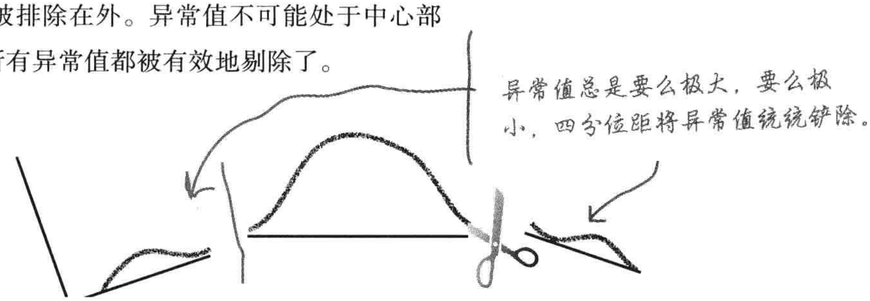
四分位距的优点是：与全距相比，较少受到异常值的影响。

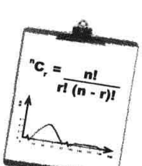
上四分位数和下四分位数所在的位置造成了这样的结果：下四分位数以下还有25%的数据，上四分位数以上还有25%的数据。也就是说，四分位距仅使用了中间50%的数据，如此将异常值弃而不用。前面已经讲过，异常值就是数据中的极大值或极小值，因此，当我们仅考虑数据中心周边的数值时，就自然而然地将异常值排除在外了。

下面再看看我们的数据。能看出四分位距如何有效地忽略异常值吗？



由于四分位距仅用了处于中心部位的50%的数据，因此，无论异常值是极大值还是极小值，均被排除在外。异常值不可能处于中心部位——这意味着，数据中的所有异常值都被有效地剔除了。





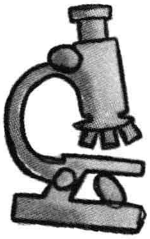
重要统计量

四分位距

即一个不易受异常值影响的“迷你距”。可通过下列方法进行计算：

上四分位数 － 下四分位数

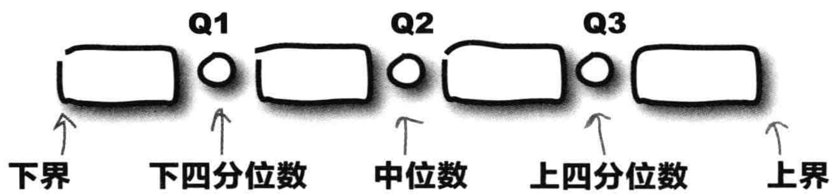
通过四分位距将异常值排除在外的意义是：得到一种对几个数据集进行比较且比较结果不会被异常值扭曲的办法。为了能算出四分位距，我们必须先算出四分位数。请翻到下一页，我们将说明如何进行计算。



剖析四分位数

求一个数据集的四分位数的过程与求中位数的过程非常相似。如果将所有数值按照升序排列，中位数就是正好位于中央的数值。如果有n个数，则中位数是位于(n+1) ÷ 2位置的数值，如果这个位置处于两个数字之间，则要取这两个数的平均值。

如果进一步将这些数据分为四份，四分位数就是处于每个分割位置的数值。最小值为下四分位数，最大值为上四分位数。



求四分位数的位置比求中位数的位置稍微棘手一点儿，因为我们需要确保所选择的数值能按正确的比例划分整批数据。不过还是有办法的：让我们从下四分位数算起。

求下四分位数的位置

1. 首先计算n ÷ 4。
2. 如果结果为整数，则下四分位数位于“n ÷ 4”这个位置和下一个位置的中间，取这两个位置上的数值的平均值，即得下四分位数。.
3. 如果“n ÷ 4”**不是**整数，则向上取整，所得结果即为下四分位数的位置。

例如，如果你有6个数，首先计算6 ÷ 4，得到1.5，向上取整得到2，这表示下四分位数的位置为2。

求上四分位数的位置

1. 首先计算3n ÷ 4。
2. 如果结果为整数，则上四分位数位于“3n ÷ 4”这个位置和下一个位置的中间，将这两个位置上的数加起来，然后除以2。
3. 如果“3n ÷ 4”不是整数，则向上取整，所得到的新数字即为上四分位数的位置。



练习

现在该实践一下你的四分位数技术了。下面是某位球员的得分：

每场比赛得分	3	6	7	10	11	13	30
频数	2	1	2	3	1	1	1



1. 这个数据集的全距是多少？

2. 下四分位数是多少？上四分位数是多少？

3. 四分位距是多少？

练习解答

下面是某位球员的得分：

每场比赛得分	3	6	7	10	11	13	30
频数	2	1	2	3	1	1	1

1. 这个数据集的全距是多少？

这个数据集的下界是3，因为3是最低得分数。上界是30，因为30是最高得分数。于是：

$$
\begin{aligned}
\text{全距} &= \text{上界} - \text{下界} \\
&= 30 - 3 \\
&= 27
\end{aligned}
$$

2. 下四分位数是多少？上四分位数是多少？

让我们先计算下四分位数。表中有11个数字，$11 \div 4 = 2.75$，将此结果向上取整可得出下四分位数的位置，因此下四分位数的位置为3，这意味着下四分位数为6。

现在让我们求出上四分位数。$3 \times 11 \div 4 = 8.25$，将此结果向上取整，得到9，即上四分位数的位置为9，这意味着上四分位数为11。

3 3 (6) 7 7 (10) 10 10 (11) 13 30

下四分位数 ↑6　中位数 ↑10　上四分位数 ↑11

3. 四分位距是多少？

四分位距等于上四分位数减下四分位数。

$$
\begin{aligned}
\text{四分位距} &= \text{上四分位数} - \text{下四分位数} \\
&= 11 - 6 \\
&= 5
\end{aligned}
$$

这个结果比全距小多了，这是因为剔除了异常值。

世上没有傻问题

问： 我明白均值、中位数、众数都很有用，可我为什么需要知道数据的分布情况呢？

答： 平均数仅能指出数据的一个方面，可以据此得知数据的中心，仅此而已，尽管很有用，但往往不够。除了平均数，还要用其他方法概括数据。

问： 这么说，中位数与四分位距是一样的喽？

答： 不对。中位数是数据的中间值，而四分位距则是50%中间数值形成的一个范围。

问： 四分位数方法有何重要意义？这似乎是一种十分繁琐的计算范围的方法。

答： 使用全距量度数据分布情况会存在一个问题：全距非常容易受异常值影响。全距能让你知道数据上界与下界之间的差值，但只要掺入一个异常值，结果就会天差地别。

解决问题的办法是：只关注居于数据中央的50%的数据，这样做能够排除异常值的干扰。这意味着要算出四分位数，并用到四分位距。因此，尽管求四分位数比求上、下界繁琐，却仍有无可置疑的优点。

问： 我总是应该用四分位距量度数据的分布情况吗？

答： 在大部分情况下，四分位距都比全距更有意义，但归根结底取决于你真正需要的信息。还有其他一些方法可以量度数据的分布情况，你可能也想考虑这些方法，我们随后将会介绍这些方法。

问： 我会不会只想看看某个四分位数，而不想看全距或四分位距？

答： 有可能。例如，你可能会有兴趣知道较大值的情况，因此你会只想看看数据集的上四分之一数据，这时你将上四分位数作为分割点。

问： 我会不会想将数据分割为比四分之一数据块更小的数据块？假如把数据分割为10份，而不是4份，结果如何？

答： 会，有时候你会想这么做。请翻开下一页，我们将具体介绍……

要点

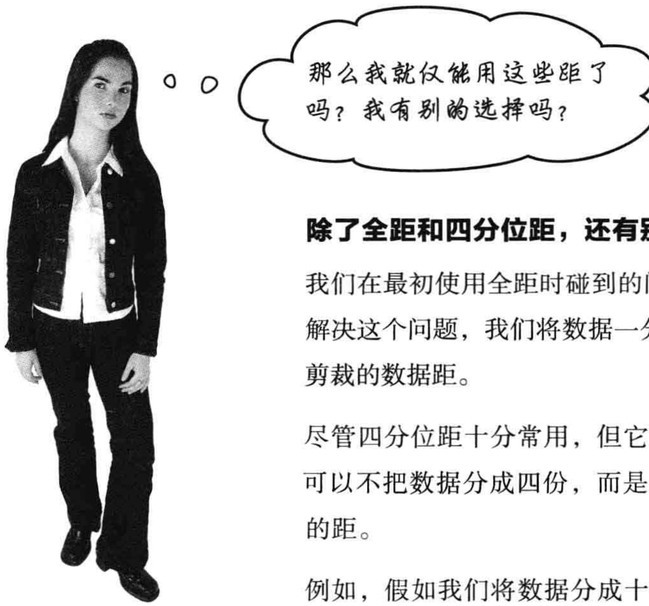
- 数据的**上、下界**即数据集中的最大值和最小值。
- **全距**是量度数据分散程度的简单方法。计算方法为：

  全距 = 上界 − 下界
- 全距很容易受异常值影响。
- 相比全距，四分位距较不易受异常值影响。
- **四分位数**即将数据分割为四等分的几个数值。最大的四分位数称为上四分位数，最小的四分位数称为下四分位数。中间的四分位数即中位数。
- **四分位距**即50%中间数值形成的一个间距。计算方法为：

  上四分位数 − 下四分位数

我们并不局限于使用四分位数

前面讲过如何通过全距和四分位距量度一批数据的数值分散情况，全距是最大值和最小值之间的差值，而四分位距则关注数据中间部位的50%数值。



除了全距和四分位距，还有别的距可供我们使用。

我们在最初使用全距时碰到的问题是：全距极易受异常值影响。为了解决这个问题，我们将数据一分为四，然后用四分位距形成一个经过剪裁的数据距。

尽管四分位距十分常用，但它并不是构建迷你距的唯一方法。我们可以不把数据分成四份，而是分为其他的份数，以此形成我们需要的距。

例如，假如我们将数据分成十份，而不是四份，使得每一个数据块包含10%的数据。于是我们就会得到如下结果：

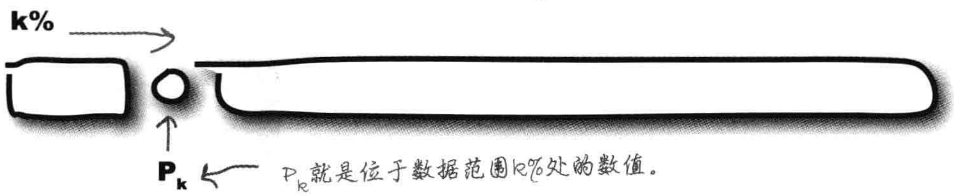


如果你将一批数据按百分比进行分割，则起分割作用的数值被称为**百分位数**。在上例中，我们的数据被分成10份，因此起分割作用的数值被称为**十分位数**。

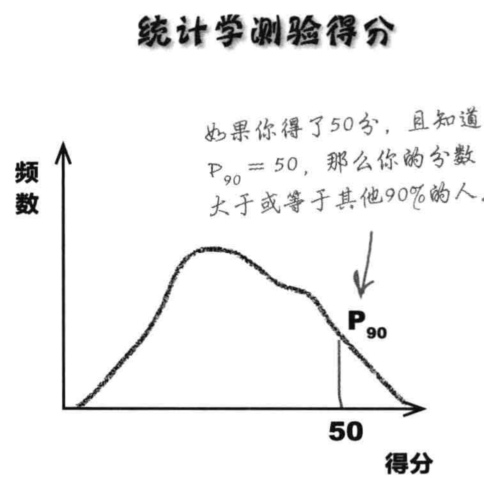
我们可以用百分位数构建一个新的距，称为**百分位距**。

什么是百分位数？

四分位数是将数据一分为四的数值，同理，百分位数是将数据一分为百的数值。每个百分位数按照它所分割出来的数据的百分比进行命名，因此，第十百分位数就是位于数据范围10%处的数值。通常，第k百分位数就是位于数据范围k%处的数值，常用P_k表示。



四分位数其实也是一种百分位数。下四分位数即P_{25},上四分位数即P_{75}。中位数即P_{50}。



百分位数用途

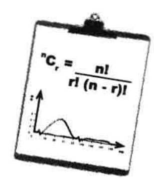
尽管百分位距不太常用，但百分位数本身却对于划分名次、排行很有用。你可以通过百分位数确定某个数值相对于其他数值的高低。例如，假定你听说自己在统计学测验中得了50分，仅看这个数字本身，你无法知道自己和别人相比是好还是坏。可如果有人告诉你这次测验的第90百分位数是50分，那么你就知道，你的分数高于或等于其他90%的人的分数。

求百分位数

求百分位数的方法与求四分位数的方法相似。

1. 首先将所有数值按升序排序。
2. 为了求出n个数字的第k百分位数的位置，先计算$k\left(\frac{n}{100}\right)$。
3. 如果结果为整数，则百分位数处于第$k\left(\frac{n}{100}\right)$位和下一位数之间。取这两个位置上的数字的平均值，得出百分位数。
4. 如果$k\left(\frac{n}{100}\right)$不是整数，则将其向上取整，结果即百分位数的位置。

例如，如果你有125个数，要求十分位数，则先计算$10 \times 125 \div 100$，结果为12.5。将此结果向上取整，得13，即十分位数为处于第13位的数值。



重要统计量

百分位数

第k百分位数即位于数据范围k%处的数值，记为：

P_k

用箱线图绘制各种“距”

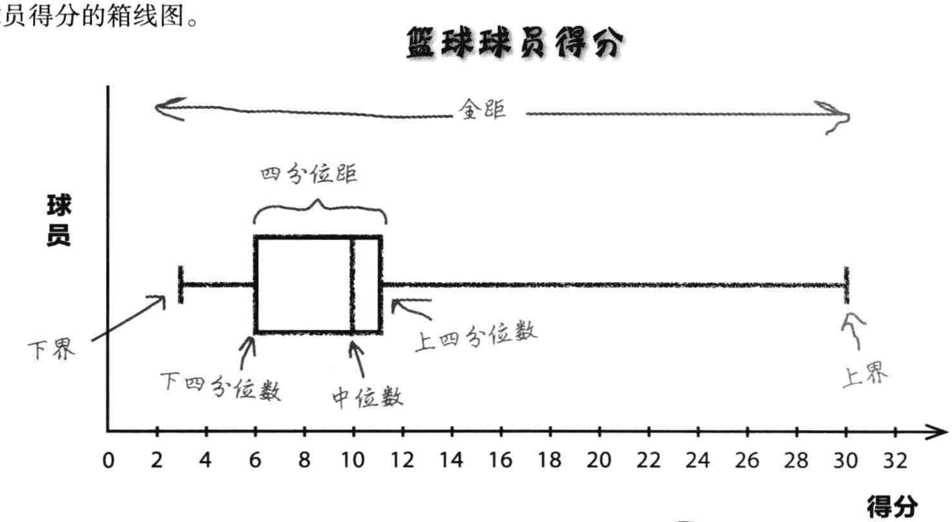
我们已经滔滔不绝地讲过各种距，如果能用直观的方法比较不同数据集的距，将会大有裨益。有一种图形专门用来显示各种各样的距，这就是**箱线图**，或者简称**箱形图**。

箱线图显示数据的全距、四分位距以及中位数。在同一张箱线图上可以比较几批数据，也就是说，箱线图是对不同数据集进行比较的极好方法。

为了创建一幅箱线图，你首先要按照标度画出一个“箱”，箱的左右两边分别代表下四分位数和上四分位数；然后，在箱中画一条线，标示出中位数；通过这个箱你能看出四分位距的宽度。随后，在箱的两边画出“线”，显示出全距的上界、下界以及宽度。以下是95页提到的球员得分的箱线图。



3 3 6 7 7 10 10 10 11 13 30



如果你的数据中有异常值，则全距会更宽。在箱线图上，一条条线的长度会随着上、下界的增长而增长。通过观察箱线图上的线，就能了解数据的偏斜程度。

如果箱线图是对称的，表示基础数据很可能也相当对称。





练习

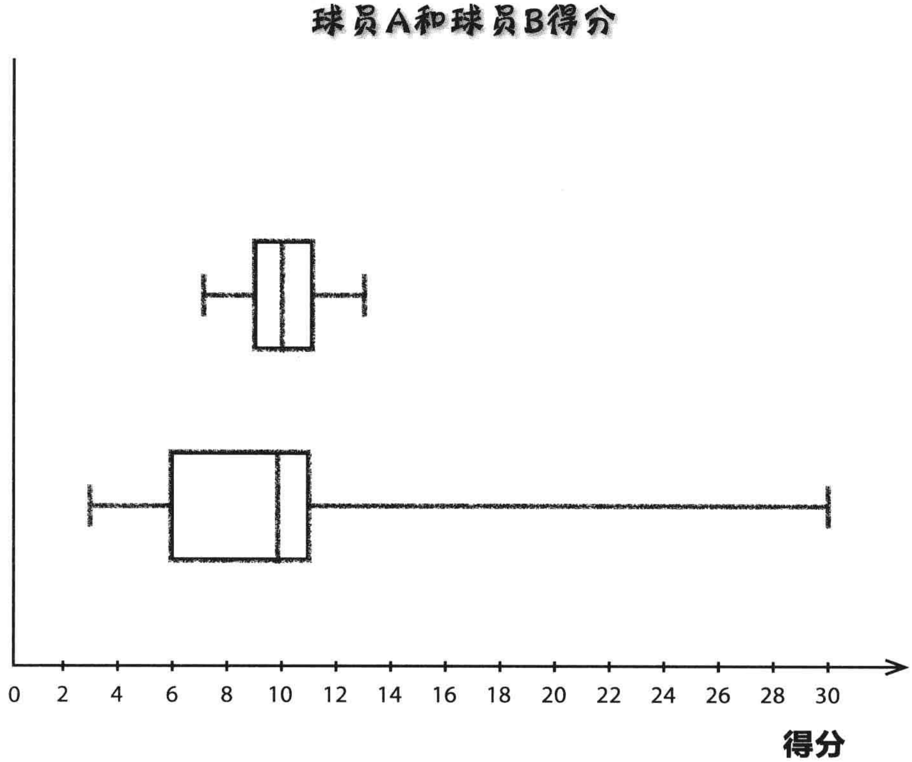
下面是另两位球员的箱线图。比较他们得分的距。如果你必须选择让球员A或球员B留在队里，你会选哪一位？为什么？



球员A



球员B



世上没有傻问题

问： 我确信我曾见过和这里的箱线图外观有所差别的箱线图。

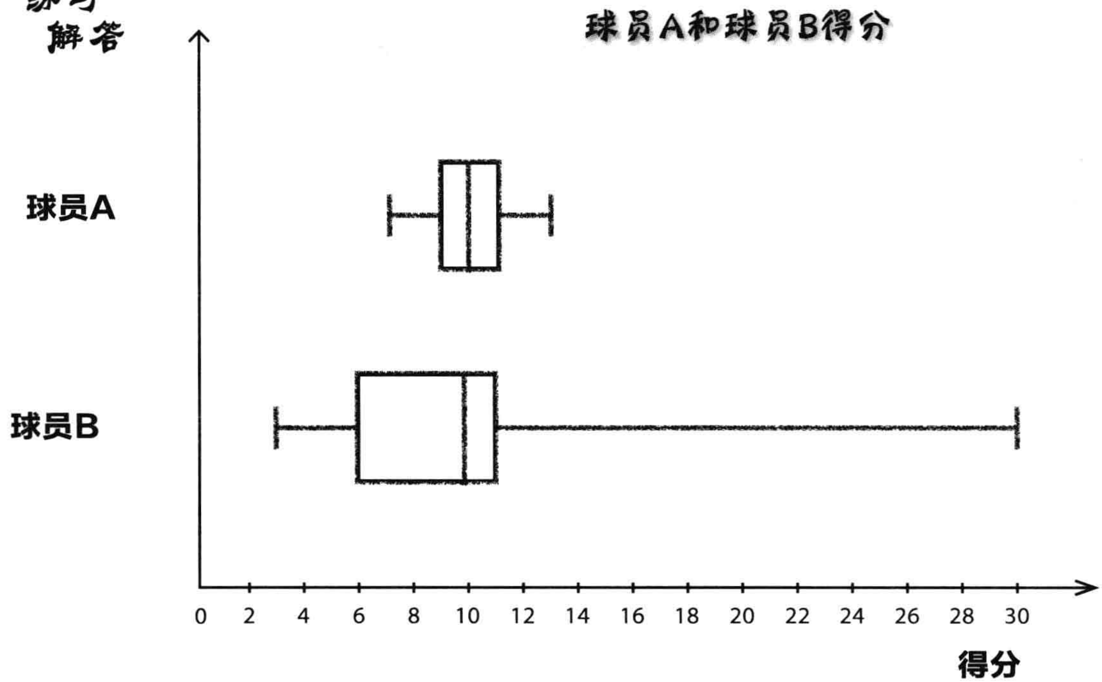
答： 箱线图确实有很多种形式。有一些形式刻意把线画短，并明确地用点或星号表示异常值，这样就很容易看出有多少异常值，以及异常值到底有多极端。另一些形式则把均值表示为点，这样你就能看出均值相对于中位数的位置。在学习统计课程的时候，查清楚有可能用到的箱线图形式是个不错的主意。

问： 那么，如果把均值表示成点，它会出现在中位数的左边还是右边？

答： 如果数据向右偏斜，则均值将位于中位数的右边，右边的线将比左边的线更长；如果数据向左偏斜，则均值将位于中位数的左边，左边的线将比右边的线更长。

练习解答

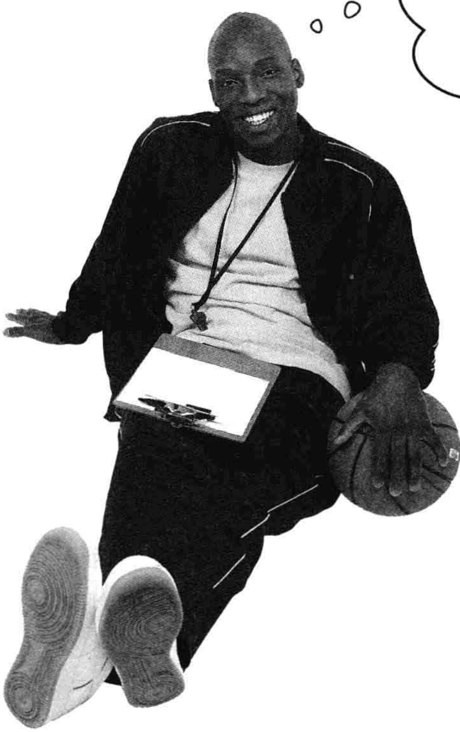
下面是另两位球员的箱线图。比较他们得分的距。如果你必须选择让球员A或球员B留在队里，你会选哪一位？为什么？



球员A的全距相对较小，他的得分中位数比球员B高一些。
球员B的全距非常大，有时候这位球员的得分比球员A高很多，但有时又低很多。
球员A发挥更稳定，通常得分高于球员B（请比较中位数和四分位距），所以，我们会选择球员A。

要点

- **百分位数**将数据一分为百。对于划分档次非常有用。
- 第k百分位数就是位于数据范围k%处的数值，用P_k表示。
- **百分位距**与四分位距相似，但百分位距是介于两个百分位数之间的距离。
- **箱线图**（或称箱形图）能在同一张图上体现多个距和四分位数，是在这方面十分有用的一种方法。“箱”显示出四分位数和四分位距的位置，“线”则显示出上、下界。箱线图能在同一张图上体现多批数据，因此非常有利于比较。



教练不仅需要比较球员得分的**全距**，他还需要以某种更为精确的方法量度大部分数值的位置所在，借此判定哪一位球员真正值得信赖，值得在比赛日委以重任。也就是说，他需要找到得分起伏最小的球员。

全距与四分位距的问题是：它们仅告诉你最大值和最小值之间的差值，却**无法告诉你**球员们得到这些最高分或最低分的频率，以及球员们得到更接近数据中心的得分的频率——而这却对教练很重要。

教练需要一支值得信赖的球员队伍，他最不想要的就是表现时好时坏，水平反复无常的队员。

为了帮助教练作出决定，我们能做点什么呢？

我们该如何更精确地量度变异性？

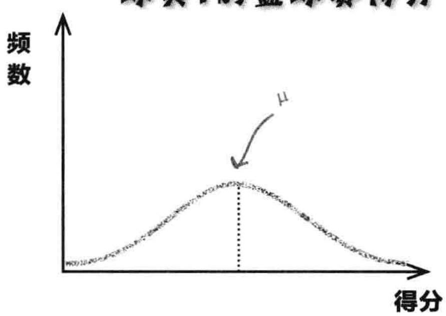
变异性比分散性更具体



我们希望量度每批得分的分散性，不止如此，还希望找到某种方法，利用所得到的分散性看出球员的稳定程度，也可以这样说：我们希望能够量度球员得分的“变异性”。

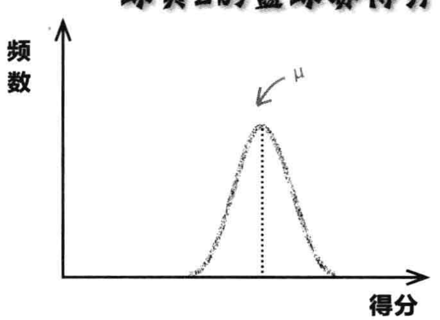
实现以上目的的一个方法是：观察每个数值与均值的距离。如果我们能够算出各个数值与均值的某种平均距离，就有办法量度变异性和分散性。结果越小，数值与均值的距离越近。下面让我们看一看。

球员1的篮球赛得分



这张图上的各个数值与均值相距甚远。如果教练把这位球员选进球队，他就不太可能预测出球员在比赛日的表现。如果这位球员在比赛日那天很顺，他或许能得极高的分；若那天很衰，他或许根本无法得高分，也就是说，球队很可能因他而败北。

球员2的篮球赛得分



这是另一位球员的得分数值，与均值的距离近得多，变化也更少。如果教练把这位球员选进球队，他会非常清楚该球员在每场比赛中可能的表现。



让我们找找答案。

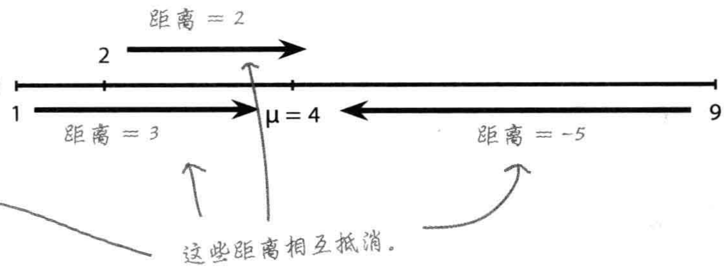
计算平均距离

假想你有三个数字：1、2、9，均值为4。如果我们求出这几个数值与均值的平均距离，结果如何？

$$\text{平均距离} = \frac{(1\text{至}\mu) + (2\text{至}\mu) + (9\text{至}\mu)}{3}$$

$$= \frac{3 + 2 + (-5)}{3}$$

$$= 0$$



各个数值与均值的平均距离总是为0 — 正负距离相互抵消。那么，我们现在该怎么办？

世上没有傻问题

问： 等式中为什么会出现－5？我会以为距离是5。距离怎么是负数呢？

答： 由于μ小于9，因此9至μ的距离为负数；1和2都小于μ，因此距离均为正数。这正是各个距离相互抵消的原因。

问： 我们不能只取正距离计算平均距离吗？

答： 这似乎很直观，但在实际应用中，统计师很少这样做。还有另一种方法能确保各个距离不相互抵消，我们很快就会讲到。这种方法能确定典型值与均值的距离远近，在统计学中广泛使用，在本书后文中的大部分章节都会出现。

问： 肯定不是所有数值的距离都会相互抵消，我们可能只是不走运而已。

答： 无论你选择哪些数值，这些数值与其均值的各个距离总是相互抵消。下面考考你：取一批数，算出均值，算出每个数值与均值的距离，然后将这些距离相加。结果次次都是0。

问： 不能用四分位距判断得分是否稳定吗？

答： 四分位距仅仅用了一部分数据来量度分散性。如果一位球员有一场比赛得分不佳，这场得分将会被剔除掉。为了实事求是地确定可靠性和稳定性，我们需要考虑所有得分。

问： 全距用上了全部得分。为什么不能用全距呢？

答： 全距仅仅在描述最大值和最小值之间的差值时才确实表现不错。如前所述，全距并不能体现数值的实际分布形态。我们需要用另一种方法进行量度。

各个数值与均值的距离正、负相抵。

我们可以用方差计算变异性……

我们要想出一个办法量度各个数值与均值的平均距离，这个办法要能防止距离与距离之间相互抵消。



让我们试着用原来的三个数字算一下。

记住：μ = 4。

$$\text{平均（距离）}^2 = \frac{(1\text{至}\mu)^2 + (2\text{至}\mu)^2 + (9\text{至}\mu)^2}{3}$$

$$= \frac{3^2 + 2^2 + (-5)^2}{3}$$

这一次是三个正数相加。

$$= \frac{9 + 4 + 25}{3}$$

$$= 12.67 \text{ (保留两位小数)}$$

这一次，各个距离没有相互抵消，我们得到了一个有意义的数。由于我们使用了各个数值与均值的距离的平方数，所有的加数都为非负数，把这些数字加起来，结果为非负数——次次如此。

这种量度数据分散情况的方法称为**方差**，是一种非常常用的描述数据分散性的方法。下面是以上等式的通用形式：

方差是数值与均值的距离的平方数的平均值。

$$\text{方差} = \frac{\sum(x-\mu)^2}{n}$$

重要统计量

方差

方差是量度数据分散性的一种方法，是数值与均值的距离的平方数的平均值。

$$\frac{\sum(x-\mu)^2}{n}$$

但标准差才是更直观的量度方法

统计师大量使用方差量度数据的分散情况。方差很有用，这是因为它用到了每一个数据，据此得出结果。可以认为方差是数值与均值的距离的平方数的平均值。



可我为什么要考虑距离的平方呢？这算不上直观。有别的办法吗？

我们真正想要的是这样一个数：能根据与均值的距离 — 而不是距离的平方指出分散性。

方差的问题是，人们恐怕难以根据距离的平方数去考虑分散性。

有一个简单的办法可对此进行修正 — 取方差的平方根，我们将此结果称为**标准差**。

让我们算出前面提到的数据集的标准差。方差为12.67，即：

$$\text{标准差} = \sqrt{12.67}$$

$$= 3.56\ (\text{保留两位小数})$$

也就是说，典型值与均值的距离是3.56。

标准差技术要诀

我们已经看出，标准差是描述典型值与均值距离的一种方法，标准差越小，数值离均值越近。标准差可能得到的最小数值为0。

像均值一样，标准差也有自己的专用符号σ，即希腊字符“西格玛”的小写（大写“西格玛”在第二章出现过：∑，表示求和）。

为了求出σ，先计算方差，然后取其平方根。

$$\sigma = \sqrt{\text{方差}}$$

↕

$$\sigma^2 = \text{方差}$$

我是标准差，要是你想量度与均值的距离，请给我来个电话。

σ



标准差访谈

本周话题：
量度标准差

Head First：嗨，标准差，见到你太好了。

标准差：很高兴见到你，Head First。

Head First：首先，我想你能不能多给我们谈谈你自己和你的工作。

标准差：我无非就是量度数据的分散性。均值很擅长让别人知道数据中心的情况，但这往往不够。有时候均值需要有人帮忙给出更完整的情况，我就是为此而来。均值体现了平均数，而我体现了数值的变异度。

Head First：恕我冒昧，我干嘛要管数据变异？这很重要吗？我肯定，只要知道一批数据的平均数就够了。

标准差：我来举个例子吧。话说你从本地餐厅定了一份快餐，当东西送到时，你发现食物一半烧焦，一半全生，这时你感受如何？

Head First：我可能会觉得不开心，觉得饿，还打算告那家餐厅。怎么了？

标准差：可是，从均值看来，你的食物是以最合适的温度烹饪的——均值显然没有体现事情的全部真相。你真正需要知道的是变异，我就是为此而来。我会根据均值体现的典型值，指出你该期望各个数值相对于这个典型值如何变化。

Head First：我想我明白了。均值给出了平均数，而你给出了分散程度。可你是怎么办到的呢？

标准差：这很简单。我不过是指出数据与均值的距离——平均而言。假定有一批数据的标准差为3cm，你可以当作这是在说：平均而言，这些数值与均值的距离是3cm。其实标准差不止包含这些信息，不过，只要顺着这样的思路去思考，你就找对方向了。

Head First：说到你的数字，标准差，你是大一点好还是小一点好？

标准差：哦，这完全取决于你要用我做什么。如果你正在生产机器零件，你会希望我小一点，这样才能确保所有的零件都一致；如果你正在研究一家大公司的工资，那么我自然会比较大。

Head First：我明白了。告诉我，你和方差有什么关系吗？

标准差：问得真好笑。方差就是另一个我——把我平方一下，我就变成方差；取方差的平方根，我就又回来了。我们两个就像是克拉克和超人，只是少件披风而已。

Head First：再问一个问题，你有没有在均值身边自惭形秽的时候？毕竟他受到的关注比你多多了。

标准差：当然没有。我们是铁哥儿们，我们相互扶持。再说，要是自惭形秽的话，会让我显得很负面——我可从来不会是负的。

Head First：标准差，感谢你的参与。

标准差：我很乐意。



练习

现在该你来显示一下标准差的实力了。请计算下列数字的均值和标准差。

1 2 3 4 5 6 7

1 2 3 4 5 6



练习解答

现在该你来显示一下标准差的实力了。请计算下列数字的均值和标准差。

1 2 3 4 5 6 7

让我们先算均值

$$\mu = \frac{1+2+3+4+5+6+7}{7} = \frac{28}{7} = 4$$

$$方差 = \frac{(1-4)^2+(2-4)^2+(3-4)^2+(4-4)^2+(5-4)^2+(6-4)^2+(7-4)^2}{7}$$

$$= \frac{3^2+2^2+1^2+0^2+(-1)^2+(-2)^2+(-3)^2}{7}$$

$$= \frac{9+4+1+0+1+4+9}{7}$$

$$= \frac{28}{7}$$

$$= 4 \qquad \sigma = \sqrt{4} = 2$$

1 2 3 4 5 6

$$\mu = \frac{1+2+3+4+5+6}{6} = \frac{21}{6} = 3.5$$

$$方差 = \frac{(1-3.5)^2+(2-3.5)^2+(3-3.5)^2+(4-3.5)^2+(5-3.5)^2+(6-3.5)^2}{6}$$

$$= \frac{2.5^2+1.5^2+0.5^2+(-0.5)^2+(-1.5)^2+(-2.5)^2}{6}$$

$$= \frac{6.25+2.25+0.25+0.25+2.25+6.25}{6}$$

$$= \frac{17.5}{6}$$

= 2.92（保留两位小数） $\sigma = \sqrt{2.92}$

= 1.71（保留两位小数）



标准差的计算可能很快就会变得错综复杂。

为了求出标准差，必须先算出方差，即求出每一个x的$(x-\mu)^2$。不过，还有一个更简单但作用相同的方差计算公式，请看下一页的内容。不过，在此之前，请你先将推导算式从奇妙池里捞出来。

奇妙池



这里藏着一个较简单的计算方差的方法，它的真面目如何？你的**任务**是将一些方程式碎片从奇妙池里捞出来，将它们放入推导过程中的空白位置。每个碎片只能用**一次**，但不需要把所有碎片都用上。**目标**：得出最后的方程式。

嘘，这里有提示。

记住：$\frac{\Sigma x}{n} = \mu$

$$\frac{\Sigma(x-\mu)^2}{n} = \frac{\Sigma(x-\mu)(x-\mu)}{n}$$

$$= \frac{\Sigma(x^2 \text{ .... } + \mu^2)}{n}$$

$$= \frac{\Sigma x^2}{n} - \frac{2\mu\,\Sigma x}{n} + \frac{\text{......}}{n}$$

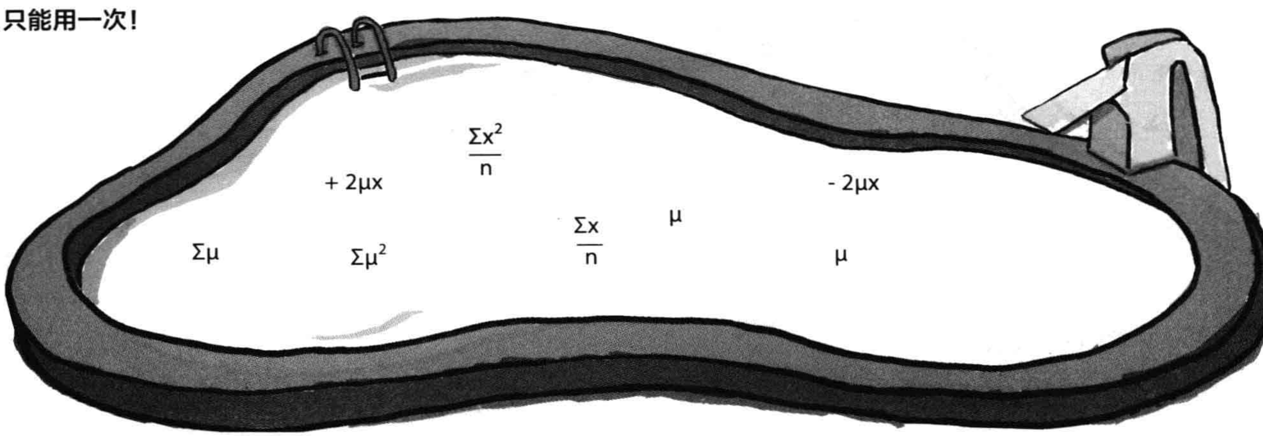
$$= \frac{\text{......}}{\text{.....}} - 2\mu \text{ ......} + \frac{n\mu^2}{n}$$

$$= \frac{\Sigma x^2}{n} - \mu_2$$

看看你能不能从这儿…

到达这儿。

注意：池中的每个算式只能用一次！



奇妙池解答



这里藏着一个较简单的计算方差的方法，它的真面目如何？你的**任务**是将一些方程式碎片从奇妙池里捞出来，将它们放入推导过程中的空白位置。每个碎片只能用**一次**，但不需要把所有碎片都用上。**目标**：得出最后的方程式。

$$\frac{\Sigma(x-\mu)^2}{n} = \frac{\Sigma(x-\mu)(x-\mu)}{n}$$

$$= \frac{\Sigma(x^2 - 2\mu x + \mu^2)}{n}$$

$$= \frac{\Sigma x^2}{n} - \frac{2\mu\,\Sigma x}{n} + \frac{\Sigma\mu^2}{n}$$

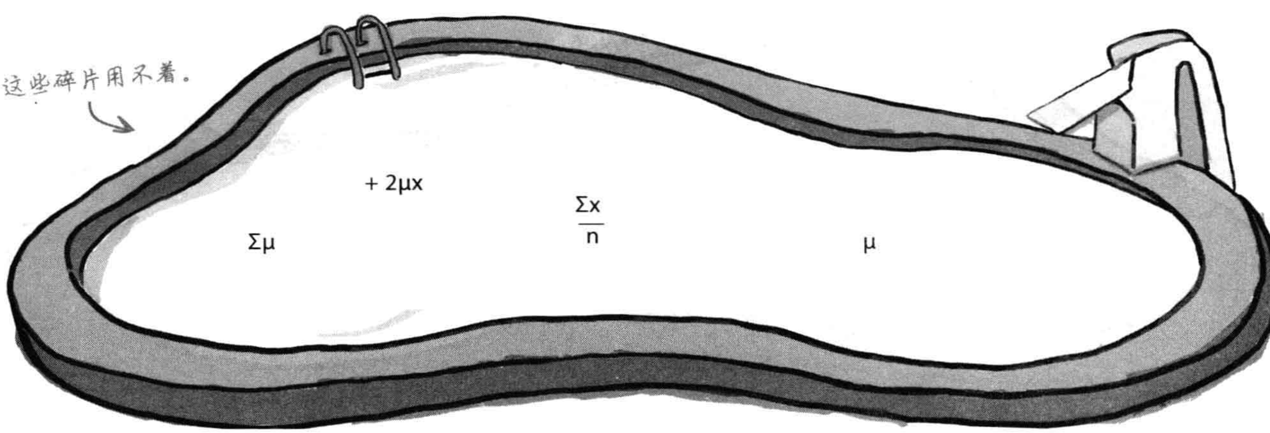
有n个。

$$= \frac{\Sigma x^2}{n} - 2\mu\,\mu + \frac{n\mu^2}{n}$$

n相互抵消。

$$= \frac{\Sigma x^2}{n} - \mu^2$$

这些碎片用不着。

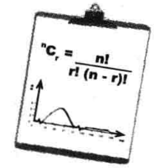


方差速算法

如前所述，标准差是量度分散性的一种方法，但为了计算标准差而进行的方差计算会迅速变得错综复杂——难就难在要计算每一个x的$(x-\mu)^2$。所处理的数据越多，就越容易出错，当μ是一个数位众多的小数时尤其如此。下面是一个能较快算出方差的方法：

$$方差 = \frac{\sum x^2}{n} - \mu^2$$

以上方法的优点是不必计算$(x-\mu)^2$。在实践中的意义是，处理起来不太麻烦，犯错误的几率也更小。



重要统计量

方差

下面是一个能较快算出方差的方法：

$$\frac{\sum x^2}{n} - \mu^2$$

世上没有傻问题

问： 那么我该用哪种形式的方差计算式呢？

答： 说到计算，第二种形式更常用，即：

$$\frac{\sum x^2}{n} - \mu^2$$

在处理小数位数众多的均值时，这种算法尤其重要。

问： 我如何用这个方差算式计算标准差？

答： 和以前一模一样，取方差的平方根即可得标准差。

问： 要是已知标准差呢？能求出方差吗？

答： 可以。标准差是方差的平方根，也就是说，方差是标准差的平方。如要通过标准差求方差，只要算出标准差的平方数即可。

问： 我发现标准差确实很费脑子。再问一遍，它是什么来着？

答： 标准差是量度分散性的一种方法，它描述了典型值与均值的距离。

如果标准差较大，意味着数值往往距离均值较远；如果标准差较小，则数值往往距离均值较近。

问： 标准差会是0吗？

答： 会。当所有数值都相同时，标准差为0。换句话说，如果每个数值与均值的距离都是0，则标准差将为0。

问： 标准差的计量单位是什么？

答： 标准差的计量单位与相应数据的单位相同。若以“厘米”进行计量，当标准差为1时，即表示在典型情况下，数值与均值相距1厘米。

问： 我肯定在你的方差计算公式中看到过除数是（n－1），而不是n，是不是哪里错了？

答： 倒是没错，不过这种形式的方差仅在处理样本时使用，本书后文谈及抽样时将详加说明。

化身教练



这里有三位球员的得分，均值都是10。你的任务就是化身为教练，算出每位球员的标准差。寻找哪一位球员是球队最靠得住的伙伴？

球员1

得分	7	9	10	11	13
频数	1	2	4	2	1

球员2

得分	7	8	9	10	11	12	13
频数	1	1	2	2	2	1	1

球员3

得分	3	6	7	10	11	13	30
频数	2	1	2	3	1	1	1



练习

星巴仕咖啡连锁店慷慨大方的首席执行官想给全体员工加薪。他拿不定主意：是直接给每个人加2,000美元呢，还是按10%的比例加。

a) 如果星巴仕每位职员都加薪2,000美元，标准差会发生什么变化?

b) 如果星巴仕每位职员都加薪10%，标准差会发生什么变化?

化身教练



这里有三位球员的得分，均值都是10。你的任务就是化身为教练，算出每位球员的标准差。寻找哪一位球员是球队最靠得住的伙伴？

球员1

得分	7	9	10	11	13
频数	1	2	4	2	1

$$方差 = \frac{7^2 + 2(9^2) + 4(10^2) + 2(11^2) + 13^2}{10} - 100$$

$$= \frac{49 + 162 + 400 + 242 + 169}{10} - 100$$

$$= 2.2$$

$$标准差 = \sqrt{2.2} = 1.48$$

球员2

得分	7	8	9	10	11	12	13
频数	1	1	2	2	2	1	1

$$方差 = \frac{7^2 + 8^2 + 2(9^2) + 2(10^2) + 2(11^2) + 12^2 + 13^2}{10} - 100$$

$$= \frac{49 + 64 + 162 + 200 + 242 + 144 + 169}{10} - 100$$

$$= 3$$

$$标准差 = \sqrt{3} = 1.73$$

球员3

得分	3	6	7	10	11	13	30
频数	2	1	2	3	1	1	1

$$方差 = \frac{2(3^2) + 6^2 + 2(7^2) + 3(10^2) + 11^2 + 13^2 + 30^2}{11} - 100$$

$$= \frac{18 + 36 + 98 + 300 + 121 + 169 + 900}{11} - 100$$

$$= 49.27$$

$$标准差 = \sqrt{49.27} = 7.02$$

球员1和球员2的标准差都很小，说明数值聚集在均值周围。而球员3的标准差为7.02，即在典型情况下，得分与均值的距离为7.02。因此，球员1是最稳定的，球员3最不稳定。



练习解答

星巴仕咖啡连锁店慷慨大方的首席执行官想给全体员工加薪。他拿不定主意：是直接给每个人加2,000美元呢，还是按10%的比例加。

a) 如果星巴仕每位职员都加薪2,000美元，标准差会发生什么变化？

标准差完全不变。实际上，数字都被拔高并向一侧移动，因此标准差不变。

$$\text{标准差} = \sqrt{\frac{\sum((x+2000)-(\mu+2000))^2}{n}}$$

$$= \sqrt{\frac{\sum(x+2000-\mu-2000)^2}{n}}$$

$$= \sqrt{\frac{\sum(x-\mu)^2}{n}}$$

= 原来的标准差

b) 如果星巴仕每位职员都加薪10%，标准差会发生什么变化？

标准差放大110%，即1.1倍。数字被拉宽了，因此标准差增大了。

$$\text{标准差} = \sqrt{\frac{\sum((1.1x)-(1.1\mu))^2}{n}}$$

$$= \sqrt{\frac{\sum 1.1^2(x-\mu)^2}{n}}$$

$$= 1.1\sqrt{\frac{\sum(x-\mu)^2}{n}}$$

= 1.1 倍原来的标准差

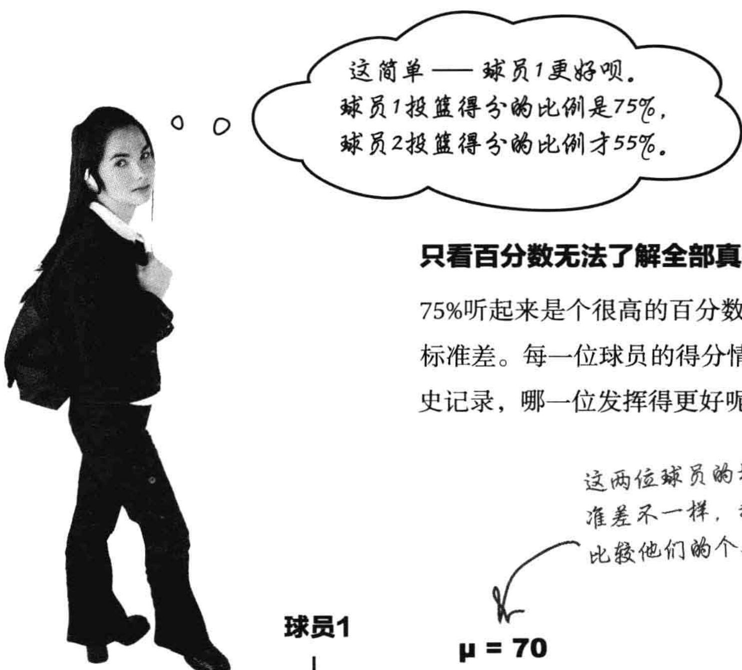
碰上需要比较基准的情况该怎么办？



我们已经讲过如何使用标准差量度一批数据的变异情况，也已经用标准差为统计邦全明星篮球队挑出了得分最稳定的球员，但标准差的用途不止于此。

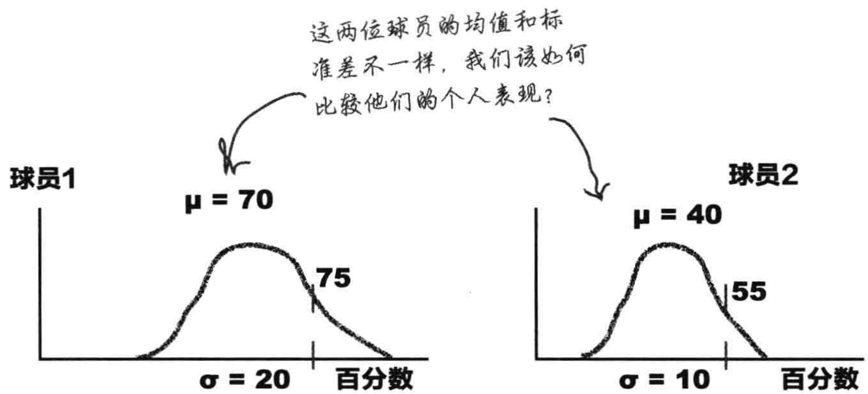
假想有两位能力不同的篮球队员：第一位投篮命中率为70%，其标准差为20%；第二位投篮命中率为40%，标准差为10%。

在某一次训练中，球员1投篮命中率是75%，球员2投篮命中率为55%。从球员本人的历史记录看来，哪一位球员的表现更好？



只看百分数无法了解全部真相。

75%听起来是个很高的百分数，但我们并不是在研究每一位球员的均值和标准差。每一位球员的得分情况都高于自己的均值，但相比球员本人的历史记录，哪一位发挥得更好呢？我们该如何对这两位球员进行比较？



这样的比较是否有可能实现？别担心，我们可以使用**标准分**(或者叫**Z分**)实现这个目的。

使用标准分比较不同数据集中的数值

使用标准分可以对不同数据集的数据进行比较，而这些不同数据集的均值和标准差各不相同 —— 标准分是对不同环境下的相关数据进行比较的一种方法。例如，你可以使用标准分比较球员相对于其本人历史记录的表现，这有点儿像私人教练的一贯做法。

通过整个数据集的均值和标准差可求出一个特定数值的标准分。标准分通常以字母“z”表示，为了求出特定数值x的标准分，可用下式进行计算：

$$z = \frac{x - \mu}{\sigma}$$

这是数值x所在的数据集的均值、标准差。

让我们算出每位球员的标准分，看看它能向我们透露什么信息。

计算标准分

让我们先算z_1，即球员1的标准分。

$$z_1 = \frac{75 - 70}{20}$$

$$= \frac{5}{20}$$

$$= 0.25$$

如上，通过用均值和标准差对得分进行标准化，球员1的得分为0.25。球员2的得分如何呢？

$$z_2 = \frac{55 - 40}{10}$$

$$= \frac{15}{10}$$

$$= 1.5$$

算得球员2的标准分为1.5，而球员1的标准分为0.25。这究竟有何意义？

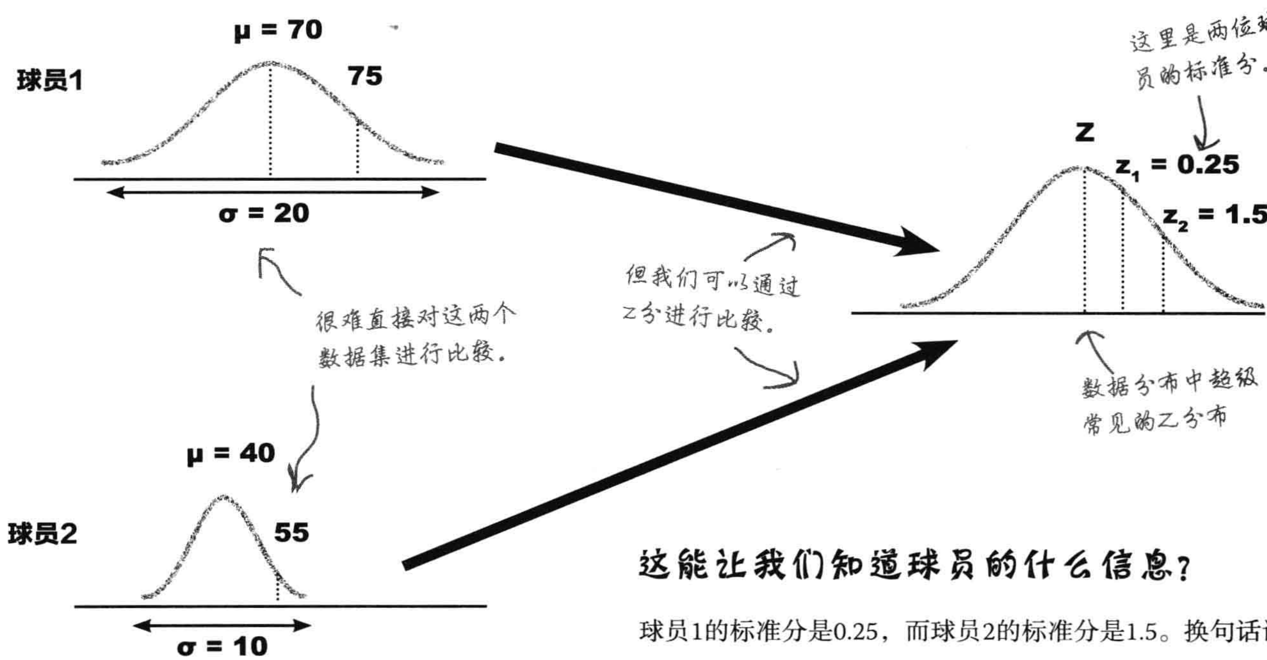
标准分释义

标准分为我们提供了一种对不同数据集的数据进行比较的办法，这些不同数据集的均值和标准差甚至都各不一样。通过这种方法，我们可以把这些数值视为**来自同一个数据集或数据分布**，从而进行比较。

而这对于我们上面提到的两位篮球队员有何意义呢？

每一位球员的投篮命中率都有不同的均值和标准差，若要比较球员们相对于自己的历史记录的表现情况，这就带来了困难。我们可以看出，在一次特定训练中，一位球员的投篮命中率高于另一位球员，我们还注意到，这两位球员的投篮命中率都比自己的平均成绩更高。难点在于要比较两位球员相对于他们本人的历史记录的表现。

标准分将每一个数据集转化为更为通用的分布形态，从而有可能进行上述比较。我们可以求出每位球员在训练中的标准分，进行转化，然后进行比较。



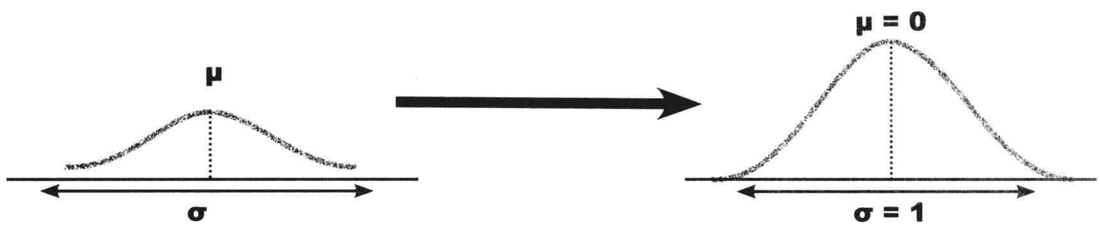
这能让我们知道球员的什么信息？

球员1的标准分是0.25，而球员2的标准分是1.5。换句话说，在将得分标准化以后，球员2的得分比球员1的得分更高。

这意味着，尽管从总体上看球员1是一位更优秀的投篮手，投篮命中率比球员2更高，但相对于本人的历史记录，却是球员2表现更好。球员2表现更好指的是……和自己比。

标准分细细看

标准分的作用是将几个数据集转换成一个理论上的新分布，这个分布的均值为0，标准差为1，这是一种可用于进行比较的通用分布。标准分将你的数据有效地转化为符合这个模型的数据，同时确保数据的基本形状不变。



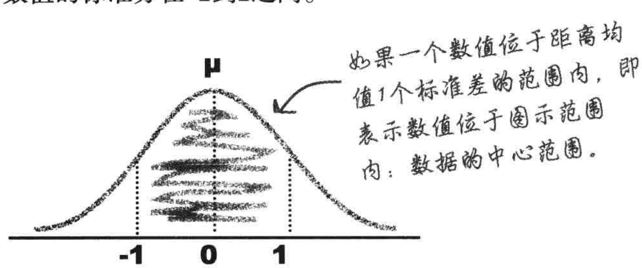
标准分可以取任意值，这些值表示相对于均值的位置。正的z分表示数值高于均值，负的z分表示数值低于均值。若z分为0，则数值等于均值本身。数值大小体现了数值与均值的距离。

距离均值若干个标准差

有时候，统计师会用**距离均值若干个标准差**表示某个特定数值的相对位置。例如，统计师可能会说某个特定值在距离均值1个标准差的范围内，这其实只不过是表示数值距离均值远近的另一种方法——它有何实际意义呢？

标准分=距离均值的标准差个数

我们已经讲过如何通过z分将数据集转化为一个均值为0、标准差为1的通用分布。如果一个数值在距离均值1个标准差的范围内，我们就知道，数值的标准分在-1到1之间。与此类似，如果一个数值在距离均值两个标准差的范围内，则数值的标准分在-2到2之间。



世上没有傻问题

问： 既然方差和标准差都能量度数据的分散程度，那么它们与全距有何区别？

答： 全距是一种极其简单的量度数据分散程度的方法，它指出最大值和最小值之间的差值，但仅此而已，你无法看出数据在这个差值范围内的聚散情况。

用方差和标准差方法量度数据的变异性和分布形态则效果好得多，因为这二者考虑了数据的聚散情况，它们关注的是典型情况下的数值与数据中心的距离。

问： 方差和标准差有何区别？我该用哪一个？

答： 标准差是方差的平方根，这说明知道其中一个就可以求出另一个。

标准差可能是最直观的方法，因为它粗略地体现了平均情况下的数值与均值的距离。

问： 标准分是如何介入以上方法的？

答： 标准分利用均值和标准差，将一个数据集中的各个数值转化为更通用的分布形态，同时确保数据的基本形状不变。

标准分是对不同数据集中的数值进行比较的一种方法——即使各个数据集的均值和标准差各不相同也能进行比较，这是一种量度相对排名的方法。

问： 标准分和异常值检测有什么关系吗？

答： 问得好！我们可以凭主观判断确定异常值，但有时候可以将异常值定义为偏离均值三个标准差的数值。

不过统计学家对此尚有分歧，因此请小心对待。

要点

- **方差**和**标准差**通过观察数值与均值的距离量度数值的分布形态。
- 方差有两种计算方法，其一：

$$\frac{\sum(x-\mu)^2}{n}$$

- 其二：

$$\frac{\sum x^2}{n}-\mu^2$$

- 标准差是方差的平方根；方差是标准差的平方。
- **标准分**（或称**z分**）是对不同数据集中的数值进行比较的一种方法，这些数据集的均值和标准差互不相同。数值x的标准分的计算方法为：

$$z=\frac{x-\mu}{\sigma}$$



填写下表。写出我们在本章讲过的各种量度分布形态的方法，说明如何进行计算，请尽量不要回头翻阅本章前面的内容。

统计量	如何计算
全距	
	上四分位数 – 下四分位数
标准差（σ）	
标准分	



填写下表。写出我们在本章见到过的各种量度分散性的方法，说明如何进行计算，请尽量不要回头翻阅本章前面的内容。

统计量	如何计算
全距	上界 - 下界
四分位距	上四分位数 - 下四分位数
标准差（ σ ）	$\sqrt{\frac{\sum(x-\mu)^2}{n}}$ $\sqrt{\frac{\sum x^2}{n}-\mu^2}$ 两种算法结果相同。
标准分	$z=\frac{x-\mu}{\sigma}$

统计邦全明星篮球队赢了联赛！

现在，整个赛季的所有比赛都结束了，统计邦全明星篮球队在联赛中排名第一。很显然，是你帮助教练选出了最适合球队的队员。

别忘了，这可都多亏标准差这位好朋友的帮助。



4 概率计算

把握机会



人生无常

瞬息之间的变化有时难以——料定。但有些事情会比其他事情更有可能发生，这就为**概率理论**提供了大显身手的舞台。通过概率能评估出现各种结果的可能性，让你**预测未来**。知悉可能出现的结果则可帮助你作出**有根据的决策**。本章将让你了解更多概率知识，学会如何掌控未来！

肥蛋大满贯

FAT DAN'S 赌场

肥蛋赌场是当地最热门的赌场，赌博游戏应有尽有 —— 轮盘、老虎机、扑克牌、二十一点……

正好你今天吉星高照，Head First实验室给了你一大堆筹码，让你去肥蛋挥霍，赢了钱全归你。想去试试？那就走吧 —— 就知道你动心了。



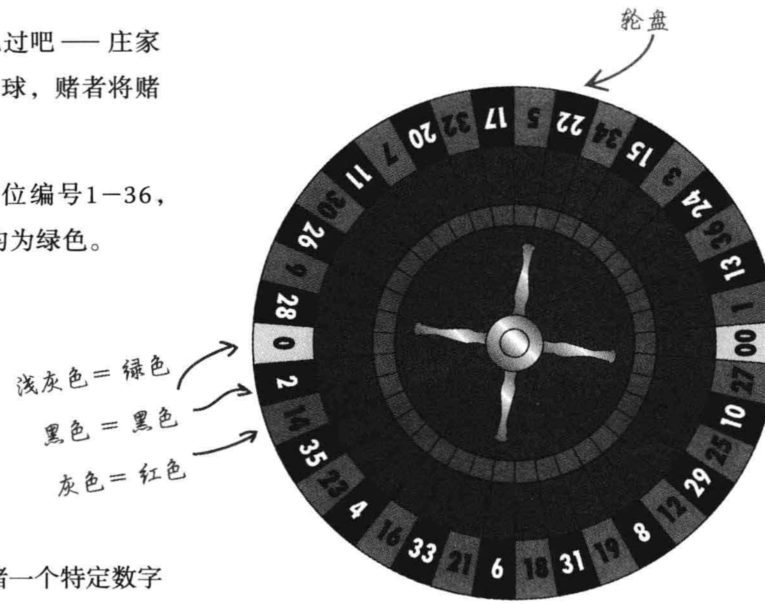


轮盘赌运转正酣，下一局正要开始，让我们看看你运气如何。

转起来吧，轮盘！

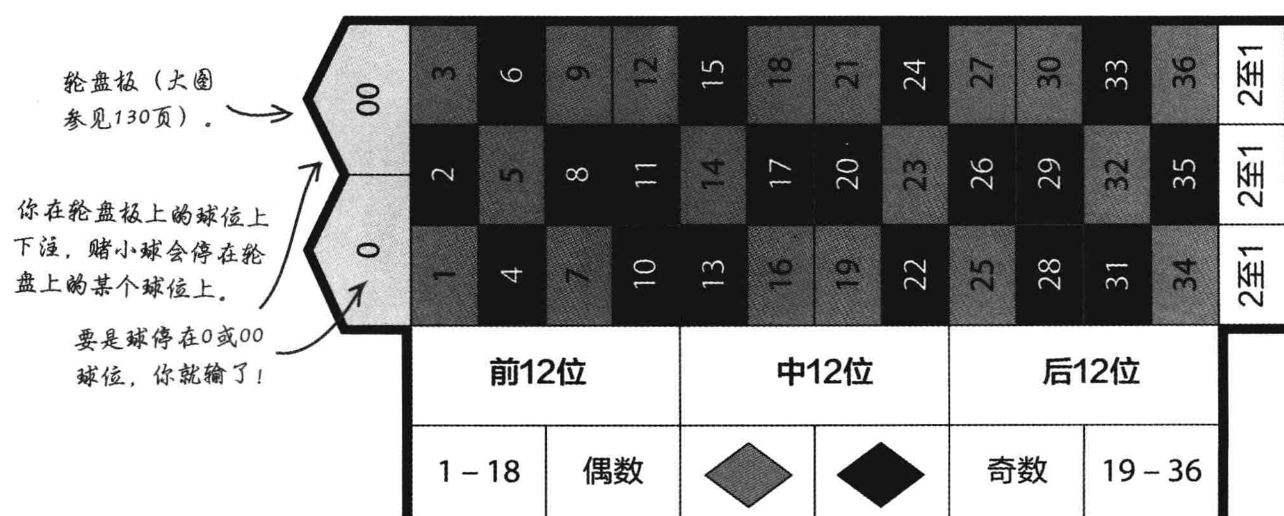
就算你没有亲自玩过轮盘赌，总在电影里见过吧——庄家转动一个轮盘，随后朝相反方向掷出一个小球，赌者将赌注押在他所料定的停球位置。

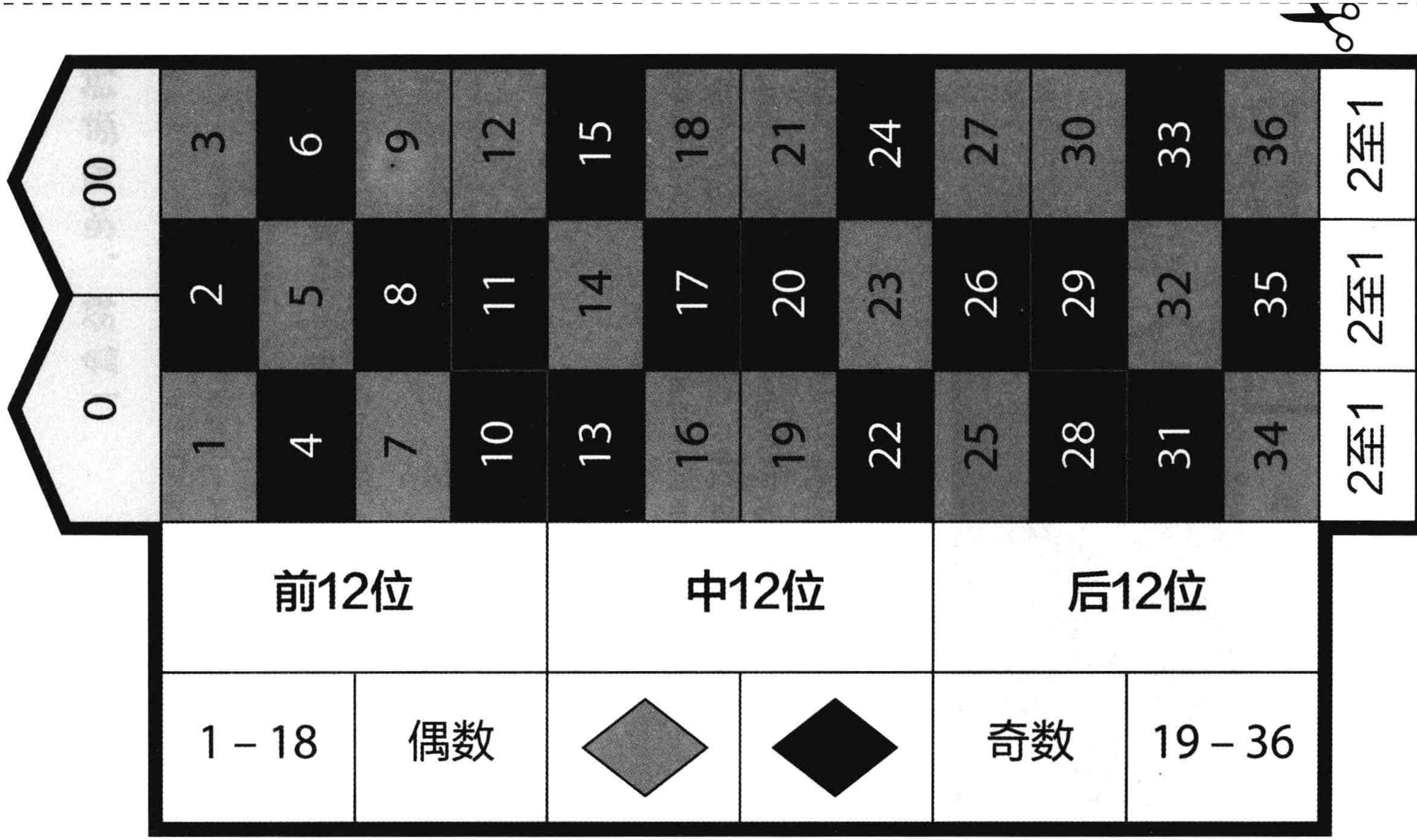
肥蛋赌场所用轮盘有38个停球位置，主球位编号1－36，颜色或黑或红；另有两个球位编号0和00，均为绿色。



轮盘赌的下注方式五花八门。例如，你可以赌一个特定数字（奇偶均可），可以赌球位颜色，开局后还会有人宣布各种其他赌法。再就是记住：如果球停在绿色球位，你就输了。

使用轮盘板可以方便地查看数字与颜色组合。





你的专用轮盘板

在本章中，你将在轮盘上大赌特赌。这里有一件称手的轮盘板，请剪下来保存好。你可以借助它计算本章要讲的概率。

小心剪刀。

下注了！

轮盘板剪好了？赌局正要开始。你料想球会停在哪里？在你的轮盘板上选择一个号码，然后下注。



正确。在下注之前，很有必要看看你有几成胜算。

也许赌中某些号码的可能性比另一些号码的可能性更大。似乎我们应该来看几个概率……

动动脑

在轮盘赌中下注前需要考虑什么？若有机会下注，你会下哪种注？为什么？

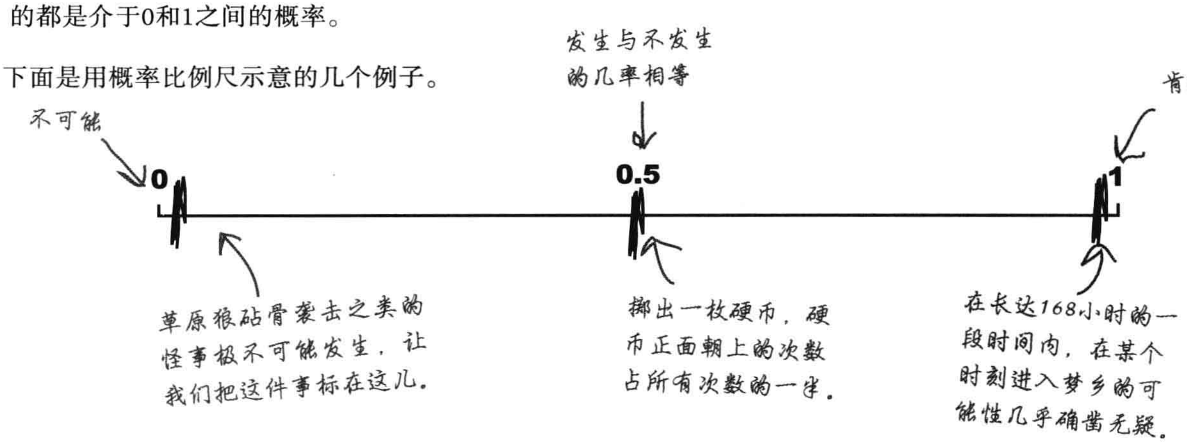
几率有多大？

当你正在思念朋友，恰好有一位朋友就给你来了电话，或者你买的彩票中了头等奖……每当这时，你会不会这样想：“那么，**这件事**的发生几率有多大？”

概率是量度某事发生几率的一种数量指标。你可以用概率衡量发生某件事的可能性（例如你在本周某一时刻会进入梦乡的可能性），或不会发生某事的可能性（例如在你徒步穿越沙漠时，草原狼企图用耳朵里的砧骨撞翻你的可能性）。统计学用“事件”一词表示有概率可言的任何事情，换句话说，**事件**就是人们能指出其发生可能性的任何事情。

概率的量度尺度是0—1。如果某件事不可能发生，则其概率为0；如果某件事肯定会发生，则其概率为1。大多数时候，你所面对的都是介于0和1之间的概率。

下面是用概率比例尺示意的几个例子。



重要统计量

众数

有概率可言的一个结果或一件事。

能看出概率与轮盘赌的关系吗？

如果你知道小球停在某个特定编号或颜色上的可能性大小，就能够判断是否该下某个赌注。若想在轮盘赌中赢钱，懂得概率是非常有用的。

动动笔

让我们来算出一个与轮盘有关的概率：小球停在数字7上的概率。
下面一步一步进行演示。

1. 观察你的轮盘板。有多少个球位可供小球停留？

2. 数字7有几个球位？

3. 为了算出“停球结果为7”的概率，用问题2的答案除以问题1的答案。结果如何？

4. 将以上概率标在下面的比例尺上。你会怎么描述“停球结果为7”这件事的可能性大小？

0　　　　　　　　0.5　　　　　　　　1

动动笔解答

你必须算出一个与轮盘有关的概率：小球停在数字7上的概率。下面我们一步一步进行演示。

1. 观察你的轮盘板。有多少个球位可供小球停留？

有38个球位 ← 别忘了：像停在其他36个球位上一样，小球还可能停在0和00球位上。

2. 数字7有几个球位？

只有1个

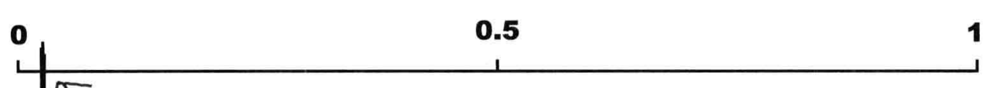
3. 为了算出“停球结果为7”的概率，用问题2的答案除以问题1的答案。结果如何？

$$“停球结果为7”的概率 = \frac{1}{38}$$

$$= 0.026$$

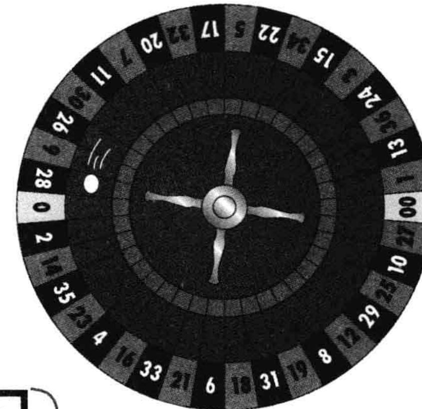
← 我们的答案，保留三位小数。

4. 将以上概率标在下面的比例尺上。你会怎么描述“停球结果为7”这件事的可能性大小？



“停球结果为7”的概率为0.026，因此是在这个位置。这个结果并非不可能，但可能性也不大。

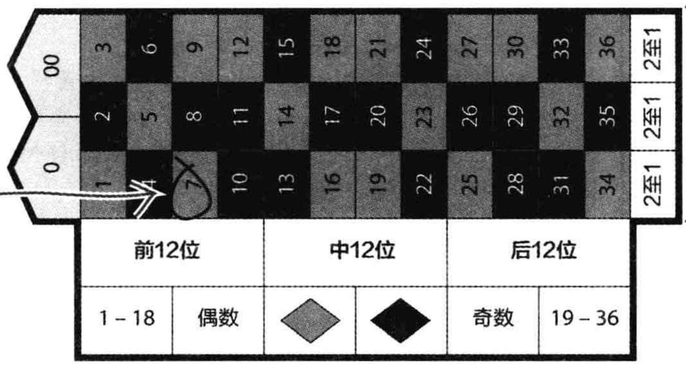
求解轮盘概率



让我们好好看看这个概率是怎么计算出来的。

下面是转动轮盘可能得到的所有结果。我们真正感兴趣的是押中赌注——即，球落在数字7上。

我们真正感兴趣的只有小球停在7的事件的概率。



这里是所有可能出现的结果，小球可能会停在其中任意一个球位上。

为了求出押中赌注的概率，我们用押中赌注的可能数目除以可能出现的结果的数目，如下所示：

$$\text{概率} = \frac{\text{押中赌注的可能数目}}{\text{所有可能结果的数目}}$$

"停球结果为7"的方式只有1种，球位有38个。

我们还可以用一种更通用的方法表述以上情况，对于事件A的概率：

$$P(A) = \frac{n(A)}{n(S)}$$

发生事件A的概率 → $P(A)$

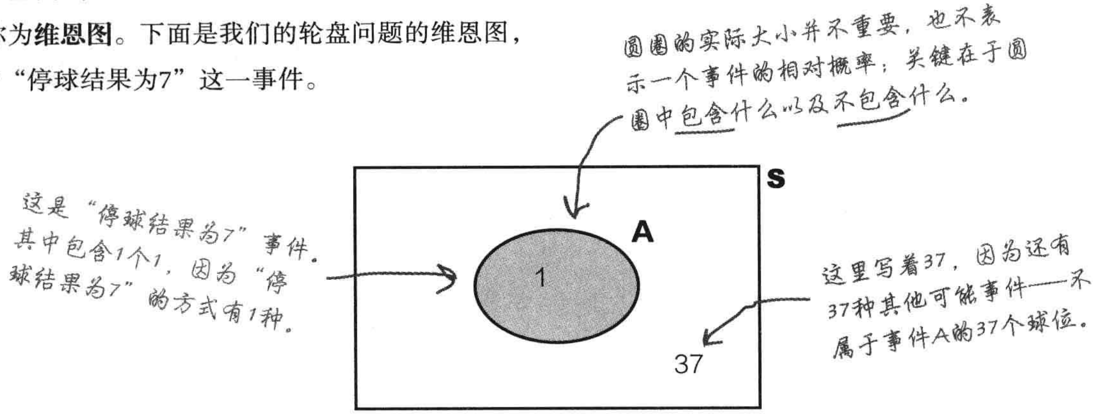
发生事件A的可能数目 → $n(A)$

所有可能结果的数目 → $n(S)$

S被称为**概率空间**，或称**样本空间**，是表示所有可能结果的一种简便表示法。可能发生的事件都是S的子集。

维恩图：概率的图形表示

概率计算有时很复杂，因此，用图形方式表示概率往往十分有用。其中有一个办法是这样的：画一个方框代表样本空间**S**，然后画几个圆圈代表各个相关事件，这种图称为**维恩图**。下面是我们的轮盘问题的维恩图，其中**A**为“停球结果为7”这一事件。



维恩图上不标出数字本身，这是十分常见的做法。你可以选择在图上标出每一事件的实际概率，以此取代数字。具体做法完全取决于你解决问题时需要用到的信息。

对立事件

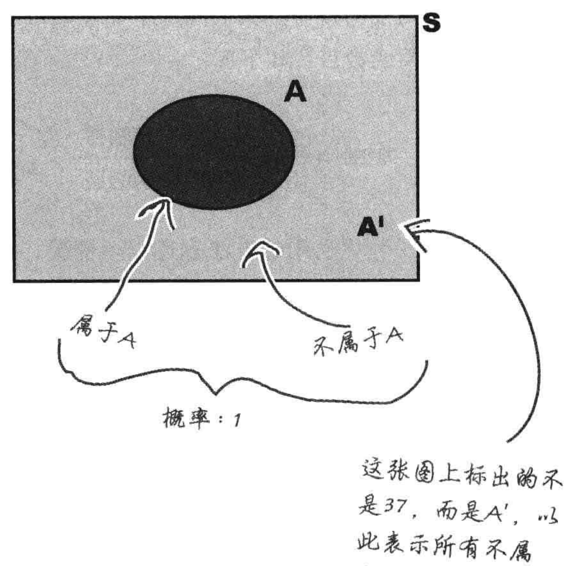
“A不发生”事件有一种简便表示方法 —— A′。A′被称为A的**对立**事件。

计算P(A′)有一种巧妙的方法。A′包含事件A所不包含的所有可能性，因此二者，即A和A′，肯定包含每一种可能发生的事件。如果某件事属于A，就不可能属于A′；如果某件事不属于A，就必定属于A′。这意味着，要是将P(A)和P(A′)加起来，结果为1。也就是说，某件事属于A或属于A′的几率为100%。于是我们得出：

$$P(A) + P(A') = 1$$

或

$$\mathbf{P(A') = 1 - P(A)}$$



化身庄家



你的任务是把自己想象成这位庄家，算出各种事件的概率。针对下列每一事件，写出获得成功的概率。

P(9)

P(绿)

P(黑)

P(38)

化身庄家解答



你的任务是把自己想象成这位庄家，算出各种事件的概率。针对下列每一事件，你应该已经获得所需结果的概率。

P(9)

“停球结果为9”的概率与“停球结果为7”的概率完全一样，因为小球落入这两个球位的几率相等。

$$概率 = \frac{1}{38}$$

$$= 0.026\ (保留三位小数)$$

P(绿)

有两个球位是绿色的，且总共有38个球位，所以：

$$概率 = \frac{2}{38}$$

$$= 0.053\ (保留三位小数)$$

P(黑)

有18个球位是黑色，且共有38个球位，所以：

$$概率 = \frac{18}{38}$$

$$= 0.474\ (保留三位小数)$$

P(38)

实际上这个事件不可能发生，因为不存在编号为38的球位。因此，这个事件的概率为0。

在讨论的几个事件中，最有可能发生的事件是小球落入一个黑色球位。

世上没有傻问题

问： 我有什么必要了解概率呢？我学的可是统计学。

答： 概率与统计学关系十分密切。大量统计知识起源于概率理论，因此懂得概率会让你的统计学技术登上一个新台阶。概率理论能帮助你进行预测，发现模式，能帮助你穿透表面上的随机性获取信息。接下来我们将会详加讲述。

问： 概率是以分数、小数还是百分数表示？

答： 可以用其中任意一种，这并不重要，只要是介于0至1之间的数值即可。

问： 我以前在集合论中看到过维恩图，这其中有关联吗？

答： 当然有。在集合论中，样本空间等于所有可能结果的集合，而可能事件则是这个集合的子集。不过，你不必为了使用维恩图计算概率而事先搞懂集合论，因为我们会在本章介绍你需要知道的各种知识。

问： 我必须画维恩图吗？我注意到你在上一个练习中并没有画。

答： 不是必须要画。但有时候，在用图形方式表示概率问题时，维恩图会是有用的工具。接下来你将看到更多有关维恩图发挥帮助作用的例子。

问： 有没有什么东西能同时存在于事件A和事件A′中？

答： 没有。A′的意思是不存在于A中的各种事物。如果某个要素存在于A中，则这个要素不可能存在于A′中。这两个事件是互斥的，因此二者不会共用任何要素。

现在该动手玩了！

一局轮盘赌即将开局。

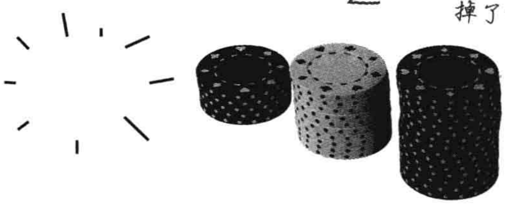
观察前一页提到的几个事件，我们将把赌注押在**最可能**发生的事件上——小球会停在黑色球位。

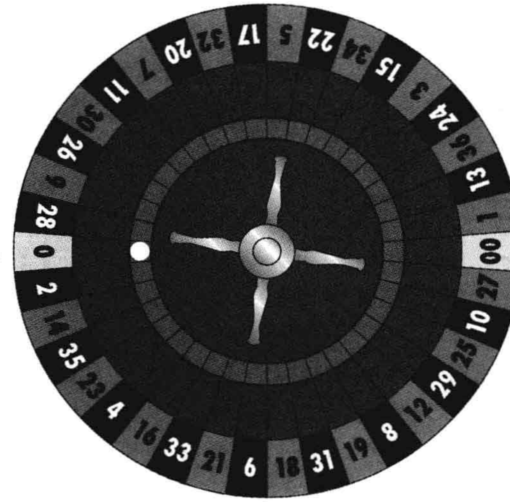


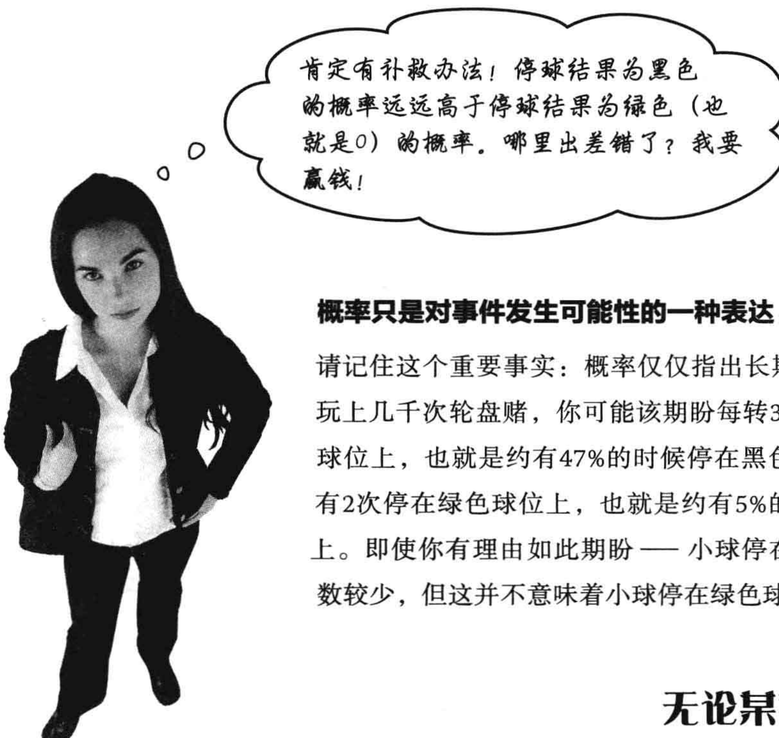
让我们看看结果如何。

赢钱的数字是……

天啊！我们算出的最可能概率事件是小球停在黑色球位上，可结果呢，它停在了绿色0球位上。你输掉了一些筹码。







概率只是对事件发生可能性的一种表达，概率并非担保。

请记住这个重要事实：概率仅仅指出长期趋势。如果你打算玩上几千次轮盘赌，你可能该期盼每转38次有18次停在黑色球位上，也就是约有47%的时候停在黑色球位上；每转38次有2次停在绿色球位上，也就是约有5%的时候停在绿色球位上。即使你有理由如此期盼 —— 小球停在绿色球位的相对次数较少，但这并不意味着小球停在绿色球位的结果不会发生。

无论某事件多么不可能发生，只要不是完全不可能发生，该事件就仍然可能发生。

让我们赌一个可能性更大的事件

让我们看一个更有可能发生的事件的概率。这次我们不押小球会停在黑色球位，而押小球会停在黑色**或**红色球位。为了算出概率，我们需要数出有多少球位是红色或黑色，用结果除以总球位数。听起来是不是很简单？





我们可以用已知的概率算出未知的概率。

看看你的轮盘板。小球只会停在三种颜色上：红色，黑色，绿色。由于我们已经算出P(绿)，于是可以用这个值求出概率，而不必数出所有的黑色和红色球位。

P(黑或红) = P(绿')

= 1 − P(绿)

= 1 − 0.053

= 0.947 (保留三位小数)

动动笔

口说无凭。数出黑色球位或红色球位的个数，用结果除以总球位数，由此算出停球结果为黑色或红色的概率。

动动笔解答

口说无凭。数出黑色球位或红色球位的个数，用结果除以总球位数，由此算出停球结果为黑色或红色的概率。

$$P(\text{黑或红}) = \frac{36}{38}$$

$$= 0.947 \text{（保留三位小数）}$$

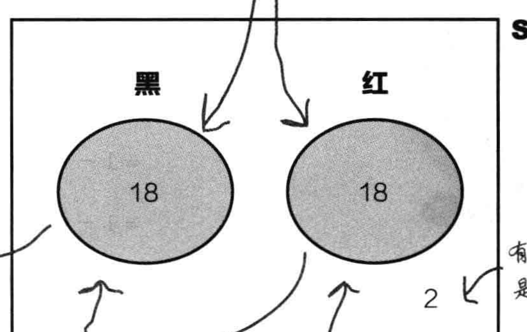
于是：$P(\text{黑或红}) = 1 - P(\text{绿})$

你还可以将几个概率相加

还有一种方法可以求出以上概率。如果已知P(黑)和P(红)，我们可以将这两个概率加起来，得出停球结果为黑色或红色的概率。让我们算算看。

球位不可能既是红色，又是黑色——这二者是“互斥事件”。

S代表样本空间，方框中包含所有的可能性。



有两个球位既不是红色也不是黑色，因此将2写在这里。

$$P(\text{黑或红}) = \frac{18 + 18}{38}$$

$$= \frac{18}{38} + \frac{18}{38}$$

$$= P(\text{黑}) + P(\text{红})$$

将几个概率相加得出的结果与将黑色或红色球位数相加再除以38的结果相同。

在本例中，将几个概率相加得出的结果与数出所有红色或黑色球位数再除以38的结果完全相同。

重要统计量

概率

如需求一个事件A的概率，算法如下：

$$P(A) = \frac{n(A)}{n(S)}$$

重要统计量

A'

A'是A的对立事件，即事件A不可能发生的事件，它的概率

$$P(A') = 1 - P(A)$$

世上没有傻问题

问： 似乎求解以上概率有三种方法，哪一种方法最好？

答： 这取决于特定情况以及你拥有的信息。

假定你拥有的关于轮盘赌的唯一信息是停球结果为绿色的概率，在这种情况下，就必须通过计算小球不停在绿色球位的概率：

$$1 - P(绿)$$

来计算要求的概率。

另一方面，如果已知P(黑)和P(红)，但颜色数目未知，则必须通过将P(黑)和 P(红)相加来计算要求的概率。

问： 这么说我不用为了计算概率而没完没了去数数了？

答： 通常不用，但还得看情况。不管怎么样，复核一下结果还是会有用的。

问： 如果某些事件发生的概率很小，人们为什么还要赌它发生呢？

答： 这和庄家所承诺的回报有很大关系。一般说来，事件的发生可能性越小，事件发生时的回报就越大。如果赌中的事件发生概率很高，那么赢的钱就不会多。人们有时会对回报率高的事件孤注一掷，即使赌赢的几率微乎其微也不惜一搏。

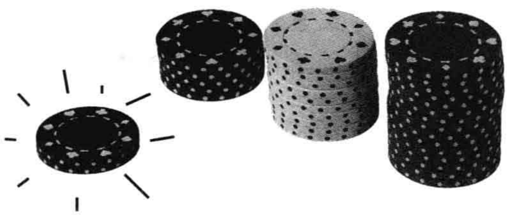
问： 像刚才那样将概率加起来总能获得正确结果吗？

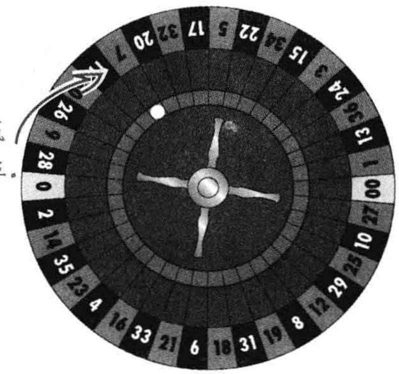
答： 请把这当作一个特例。其他情况我们将在接下来的几页中详细讲解。

你赢钱了！

这一回，小球停在红色球位上，数字是7，因此你赢了一些筹码。

这一回，你选中了一个赢钱的球位：一个红色球位。

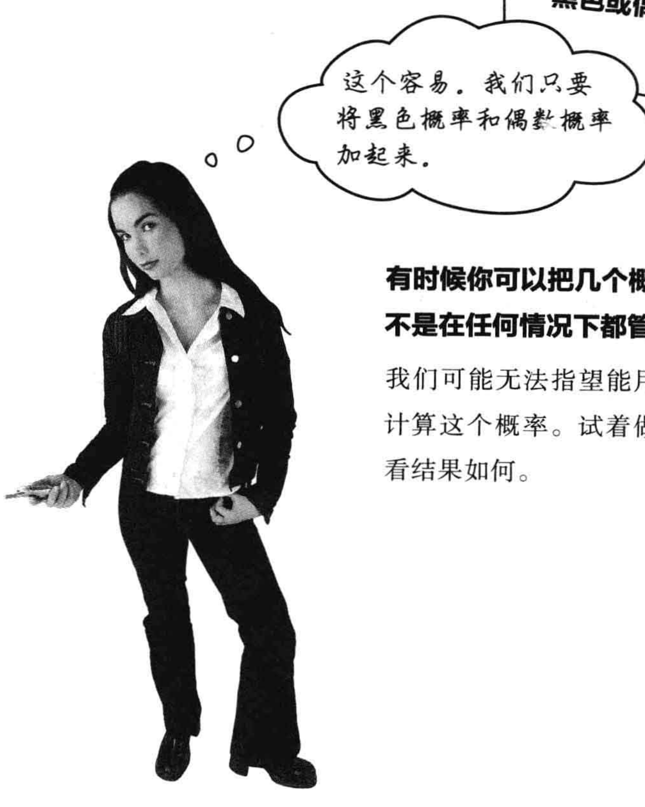




再赌一局

既然你已经掌握了计算概率的窍门，那就让我们试着算点别的东西吧：小球停在黑色或偶数球位上的概率是多少？





有时候你可以把几个概率加起来，但这一招并不是在任何情况下都管用。

我们可能无法指望能用和前面完全一样的算法计算这个概率。试着做一做下一页的练习，看看结果如何。

动动笔

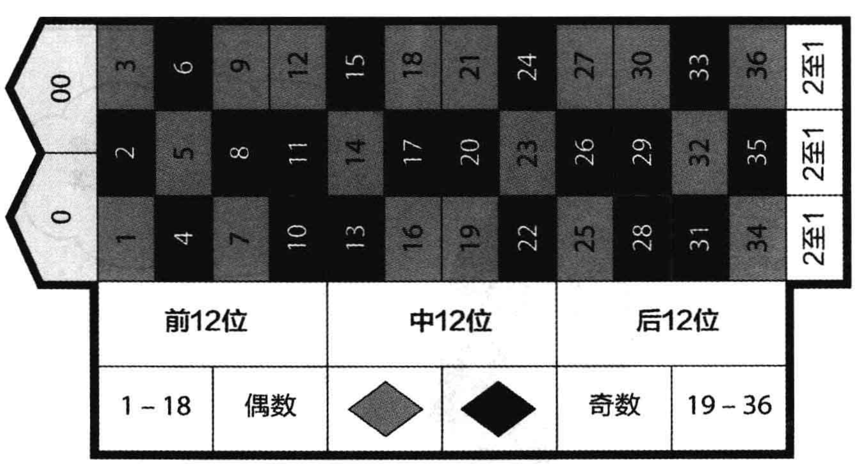
让我们求出“停球结果为黑色或偶数”的概率（假设0和00不是偶数）。

1. “停球结果为黑色”的概率是多少？

2. “停球结果为偶数”的概率是多少？

3. 将以上两个概率相加，结果如何？

4. 最后，用你的轮盘板数出所有的黑色或偶数球位，然后除以球位总数。结果如何？



动动笔解答

让我们求出“停球结果为黑色或偶数”的概率（假设0和00不是偶数）。

1. “停球结果为黑色”的概率是多少？

   18 / 38 = 0.474

2. “停球结果为偶数”的概率是多少？

   18 / 38 = 0.474

3. 将以上两个概率相加，结果如何？

   0.947

4. 最后，用你的轮盘板数出所有的黑色或偶数球位，然后除以球位总数。结果如何？

   26 / 38 = 0.684

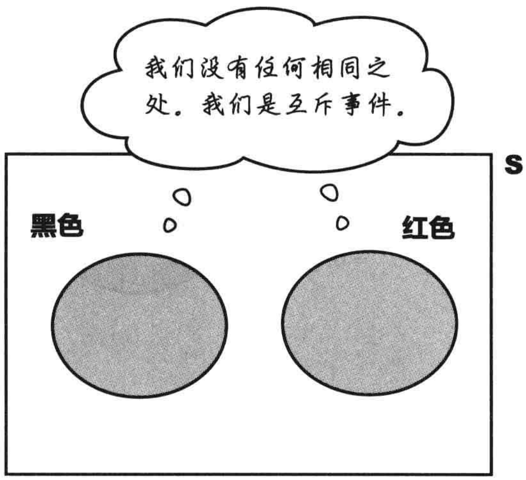
哎呀！答案不一样



让我们仔细看看……

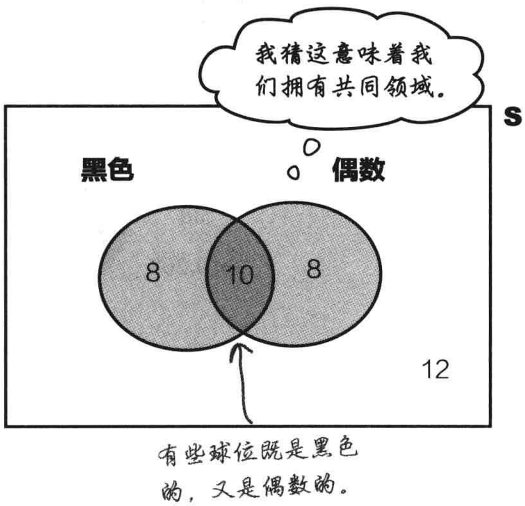
互斥事件与相交事件

在我们计算小球停在黑色球位或红色球位上的概率的时候，所面对的是两个互斥事件——小球停在黑色球位上、小球停在红色球位上。由于小球不可能既停在黑色球位上，又停在红色球位上，因此这两个事件是**互斥的**。



如果两个事件是互斥事件，则只有其中一个事件会发生。

黑色球位事件和偶数球位事件又是怎样的关系呢？这一次，这两个事件不互斥，小球有可能既停在黑色球位上，又停在偶数球位上。这两个事件是**相交事件**。



如果两个事件相交，则这两个事件有可能同时发生。

动动脑

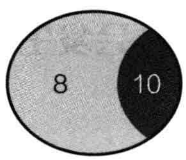
你觉得这种相交状况会对概率有何影响？

交集带来的问题

“停球结果为黑色或偶数”的计算结果之所以出现差异，是因为我们将“黑色兼偶数”球位算了两次。下面是具体分析。

首先，我们求出“停球结果为黑色”的概率以及“停球结果为偶数”的概率。

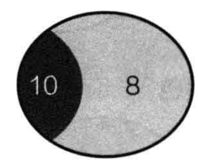
黑色



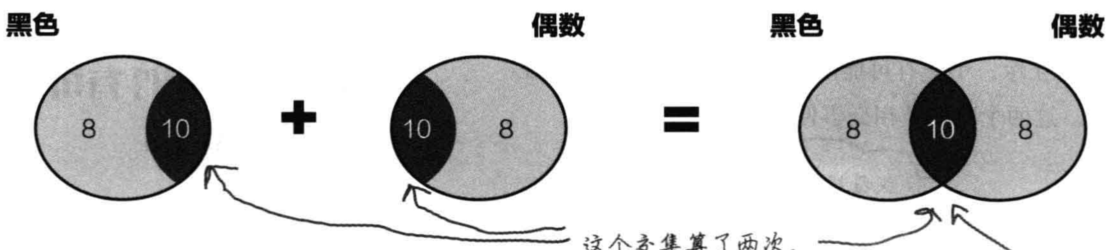
$$P(黑) = \frac{18}{38} = 0.474$$

偶数

$$P(偶) = \frac{18}{38} = 0.474$$



当将两个概率相加时，我们将停球结果为“黑色兼偶数”的概率算了两次。



$$P(黑 \cap 偶) = \frac{10}{38} = 0.263$$



为了得出正确的答案，须减去停球结果为“黑色兼偶数”的概率。得到：

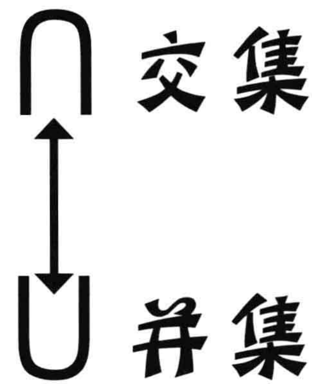
P(黑或偶) = P(黑) + P(偶) − P(黑兼偶)

这一部分只能算一次，因此让我们减去其中重复的部分。

下面可以代入我们前面算出的值，以便求出P(黑或偶)：

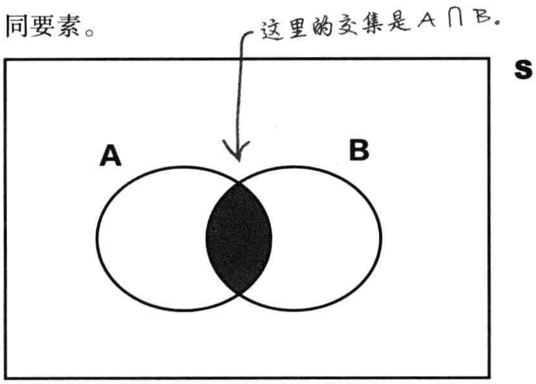
P(黑或偶) = 18/38 + 18/38 − 10/38 = 26/38 = 0.684

更多表示法



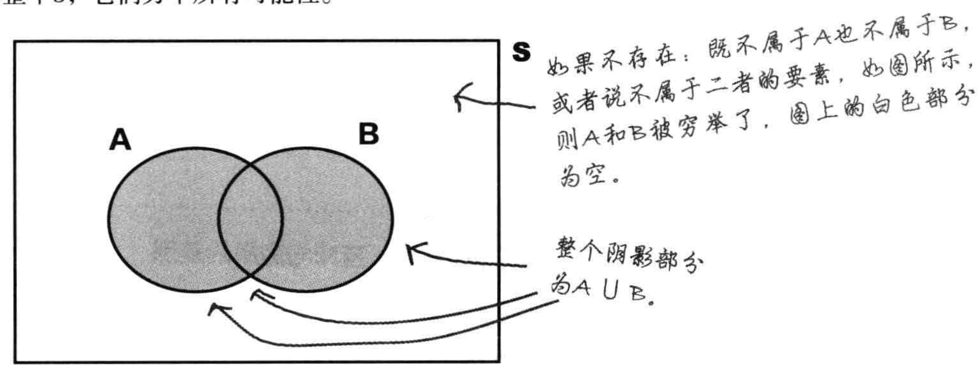
还有一种更通用的表示法，其中使用了更多简便的数学符号。

首先，我们可以用A∩B表示“A与B的交集”，你可以把这个符号理解为“与”，它求出不同事件的共同要素。



另一方面，A∪B则表示“A与B的并集”，它包含属于A及B的所有要素，你可以把这个符号理解为“或”。

如果P(A∪B)=1，则我们说A与B**穷举**。它们一起形成整个S，它们穷举所有可能性。



动动笔

我们在上一页得出

P(黑或偶) = P(黑) + P(偶) − P(黑和偶)

请用∩和∪符号表示上式。

动动笔解答

P(黑或偶) = P(黑) + P(偶) - P(黑和偶)

请用∩和∪符号表示上式。

P(A或B) →

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

← P(A与B)



实际上并无太大差别。

互斥事件之间并无相同要素。如果你有两个互斥事件，则“A交B”的计算结果其实为0—— 即$P(A \cap B) = 0$。让我们再看看黑色球位或红色球位的例子。对于这个赌注，轮盘上的“停球结果为红色球位”与“停球结果为黑色球位”这二者是互斥的，因为球位不可能既是红色又是黑色，即P(黑 ∩ 红) = 0，因此表示这一部分的等式就不见了。



互斥与穷举的差别

如果事件A与事件B为互斥事件，则

$$P(A \cap B) = 0$$

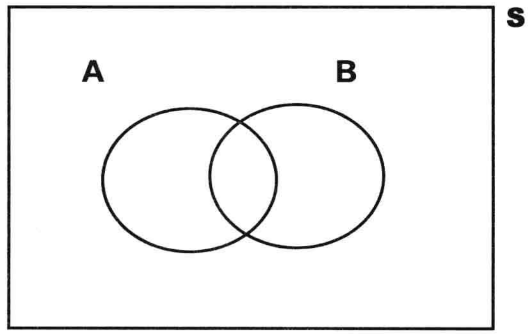
如果事件A与事件B为穷举事件，则

$$P(A \cup B) = 1$$

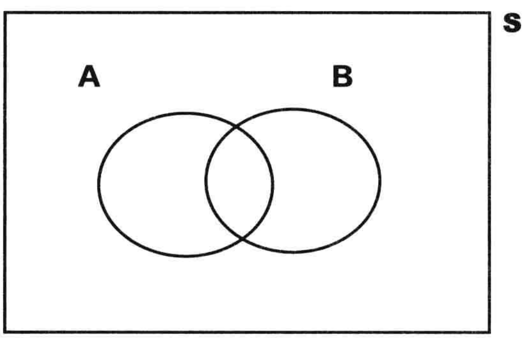


化身概率

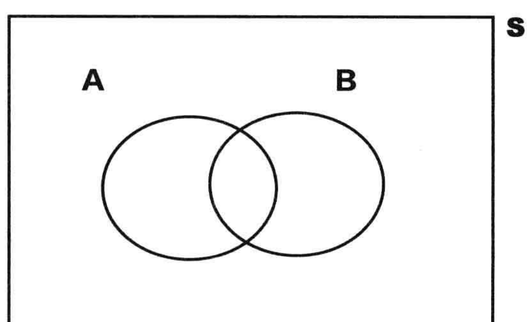
你的任务是扮演概率，把维恩图上代表下列概率的部位涂上阴影。



$P(A \cap B) + P(A \cap B')$



$P(A' \cap B')$

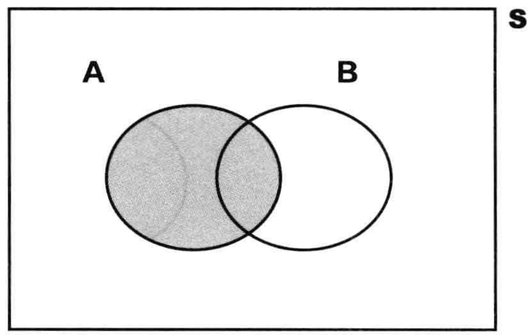


$P(A \cup B) - P(B)$

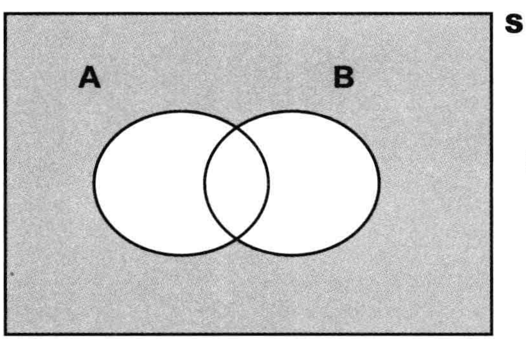


化身概率解答

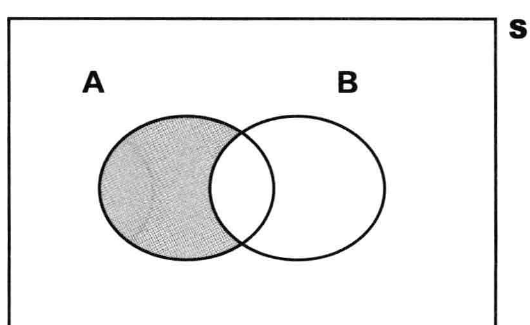
把维恩图上代表下列概率的部位涂上阴影。



P(A ∩ B) + P(A ∩ B')



P(A' ∩ B')



P(A ∪ B) - P(B)

练习

Head First健康俱乐部有50位运动爱好者接受了调查，调查问及他们是否打棒球、篮球或踢足球。结果有10位运动爱好者仅打棒球，12位仅踢足球，18位仅打篮球；6位既打棒球又打篮球，但不踢足球；4位既踢足球又打篮球，但不打棒球。

画一张维恩图代表这个概率空间。总共有几位运动爱好者打棒球？几位打篮球？几位踢足球？

以上运动花名册有没有互斥的？哪些运动是穷举的（填满概率空间）？

重要统计量

A或B

为了求出以事件A或B为结果的概率，可以使用下列算法：

$$P(A \cup B) = P(A) + P(B) - P(A \cup B)$$

$\cup$ 表示“或”

$\cap$ 表示“与”

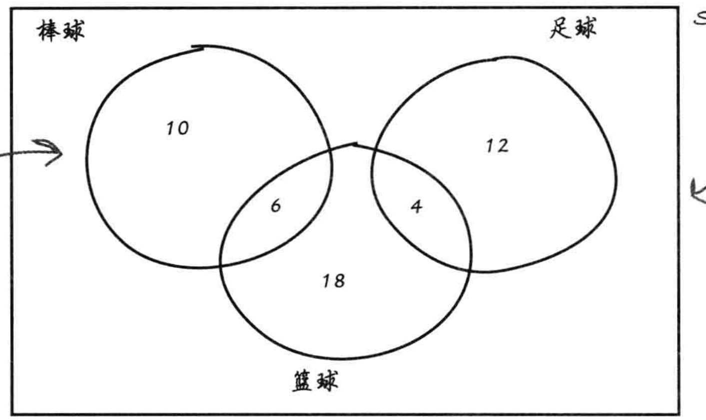
练习解答

Head First健康俱乐部有50位运动爱好者接受了调查，调查问及他们是否打棒球、篮球或踢足球。结果有10位运动爱好者仅打棒球，12位仅踢足球，18位仅打篮球；6位既打棒球又打篮球，但不踢足球；4位既踢足球又打篮球，但不打棒球。

画一张维恩图代表这个概率空间。总共有几位运动爱好者打棒球？几位打篮球？几位踢足球？

以上运动花名册有没有互斥花名册？哪些运动是穷举的（填满概率空间）？

将已知数据全部加起来，结果为50，即运动爱好者总数。



图上的信息看起来错综复杂，不过，绘制维恩图将有助于我们看清形式。

通过将各个圆圈中的数值相加，我们可以确定：棒球爱好者的总数为16，篮球爱好者的总数为28，足球爱好者的总数为16。

棒球事件和足球事件为互斥事件，没有任何人既打棒球又踢足球，因此P(棒球 ∩ 足球) = 0。

棒球事件、篮球事件和足球事件是穷举的，它们共同填满了整个概率空间，因此P(棒球 ∪ 足球 ∪ 篮球) = 1

世上没有傻问题

问： A和A′是互斥的还是穷举的？

答： 其实两样都是。A和A′不可能有任何共同要素，因此二者互斥；若将二者相加，则形成整个概率空间，因此二者穷举。

问： P(A ∩ B) + P(A ∩ B′)不就是P(A)的复杂化表示方法吗？

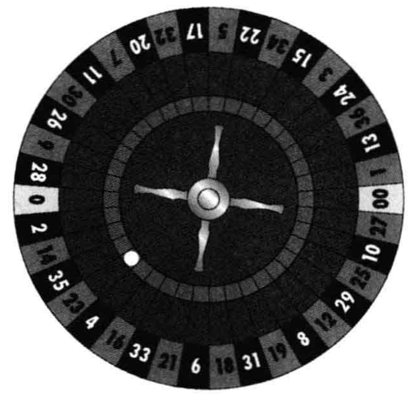
答： 是啊，正是如此。不过有时候，想出不同的方法表示同样的概率挺有用的。你并不总是能得到希望得到的信息，因此，改变一下思维方式绝对是一个优势。

问： 相交事件的数量是否受到限制？

答： 并无限制。几个事件的交集可以多用几个∩符号表示。例如，事件A、B、C的交集用A∩B∩ C表示。

有时候，求几个交集的概率很是棘手，若遇到麻烦，建议画一幅维恩图，并认真、专注地查看要将哪几个概率加起来，以及要将哪几个概率减去。

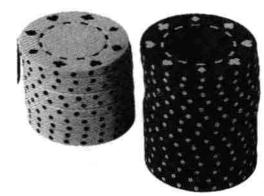
又一次倒霉的转动……

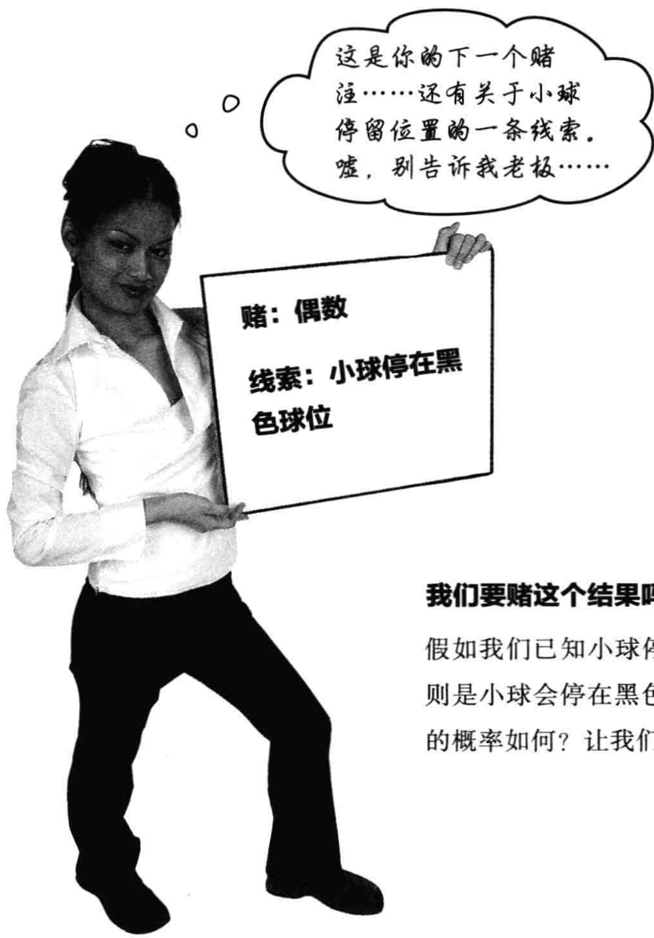


我们已知小球停在黑色或偶数球位上的概率为0.684，可倒霉的是，小球停在了23位 —— 红色，奇数。

不过另一局又要开场了

即使是我们喜欢的奇数也不能给我们带来轮盘赌上的好运。庄家决定发发善心，给我们一点点内幕消息。她将在转动轮盘后给我们一条有关小球停留位置的线索，而我们呢，将根据她的线索算出概率。





我们要赌这个结果吗？

假如我们已知小球停在黑色球位上 —— 而上一局则是小球会停在黑色或奇数上，那么结果为偶数的概率如何？让我们算一算。

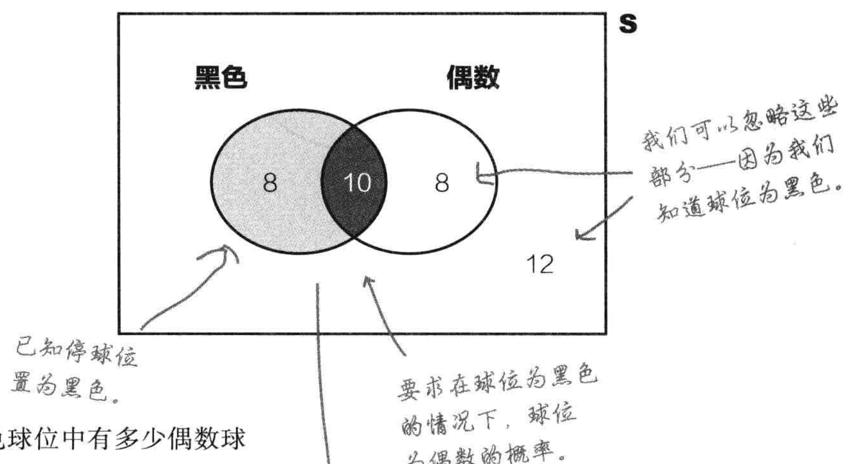
设定条件

庄家说小球停在黑色球位，那么小球同时停在偶数球位的概率是多少？



问题略有区别

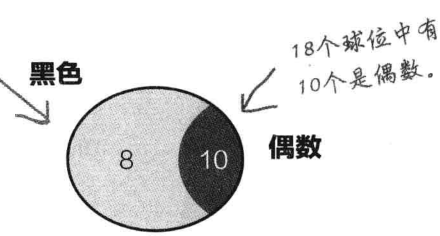
我们要算的不是“停球结果为黑色与偶数”相对于“全部可能停球位置”的概率，而是在“已知停球位置为黑色”的情况下，求“球位为偶数”的概率。



换言之，我们要求出在所有黑色球位中有多少偶数球位。在18个黑色球位中，有10个球位是偶数，所以：

$$P(\text{黑色已知条件下的偶数}) = \frac{10}{18}$$

$$= 0.556\ (\text{保留三位小数})$$



结果证明，即使得到了内幕消息，我们的胜算实际上还是较之前低。“黑色已知条件下的偶数”的概率实际上小于“停球位置为黑色或偶数”的概率。

不过，0.556这个概率仍然比50%的胜算更大，因此仍是一个不错的赌注。让我们继续。

求解条件概率

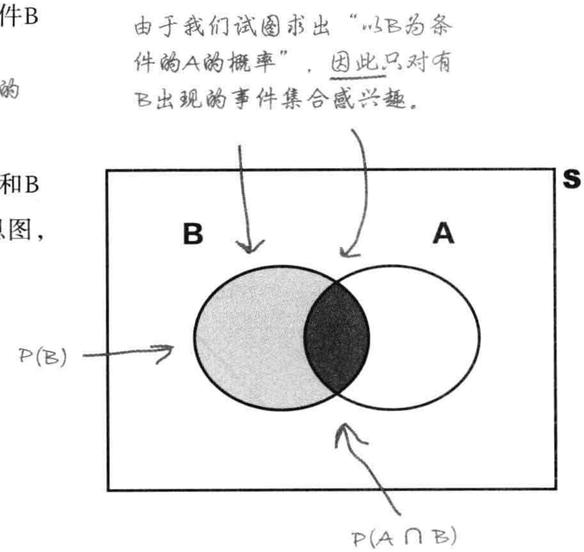


该怎么归纳这一类问题呢？首先，我们要另用一种表示法表示**条件概率**，用它来量度与其他事件的发生情况有关的某个事件的概率。

如果要表示以另一个事件的发生为条件的某个事件的发生概率，我们就用"|"符号表示"已知条件"，于是，"以事件B为已知条件的事件A的概率"就可以简写为：

$$P(A \mid B)$$

在已知B已经发生的条件下A的概率。



现在要用一种通用方法来计算P(A|B)。我们感兴趣的是A和B同时发生的次数与B发生的所有次数相除的结果。观察维恩图，得到：

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

我们将算式改变一下，以便得出求P(A ∩ B)的方法：

$$P(A \cap B) = P(A \mid B) \times P(B)$$

这还不是最终结果，P(A∩B)的另一种表示方法是P(B∩A)，即我们可以将算式写成：

$$P(B \cap A) = P(B \mid A) \times P(A)$$

也就是将A和B对调一下。



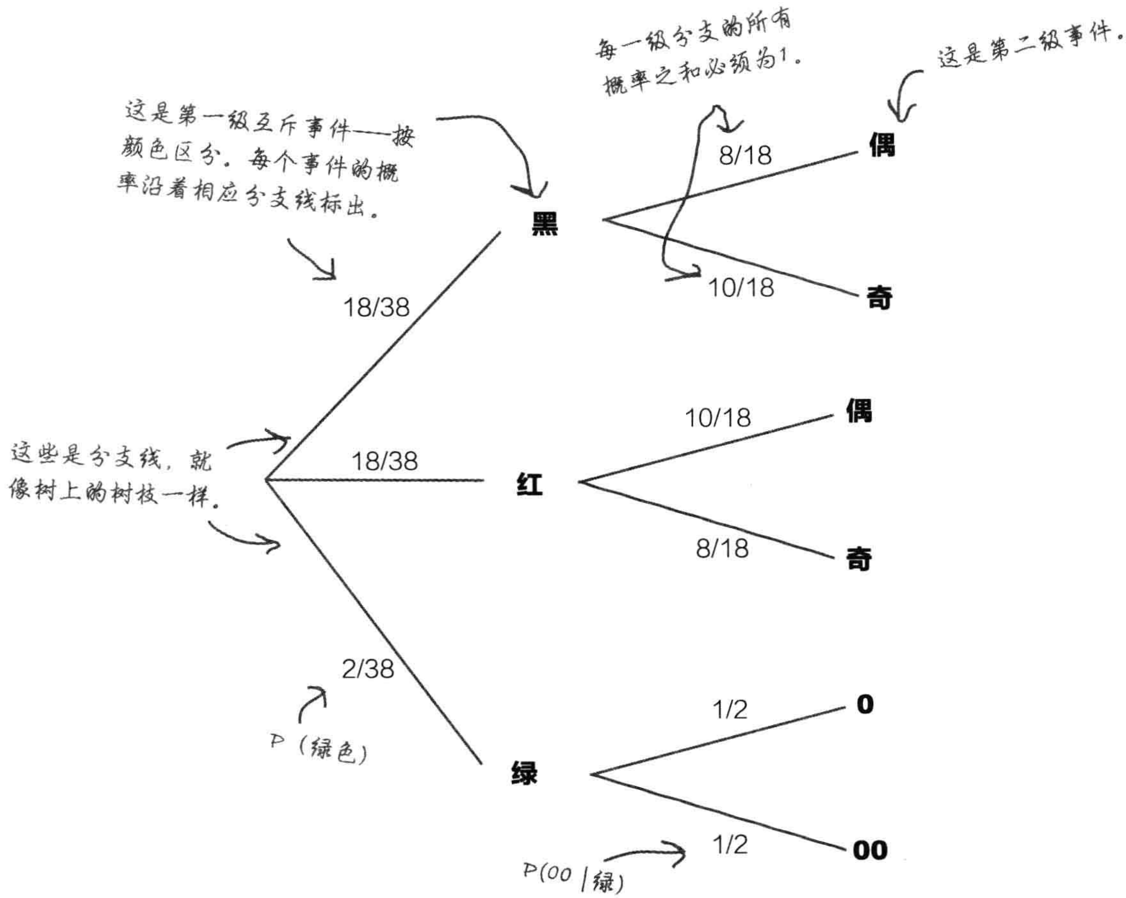
维恩图并不总是表示条件概率的最好方法。

别担心，还可以用另一种图 —— 概率树。

用概率树表示条件概率

用维恩图表示条件概率并不总是那么方便，但还有另一种图形，倒是能得心应手地处理条件概率，这就是**概率树**。

下面是关于轮盘问题的一幅概率树，其中标有以几种颜色的球位以及奇偶球位为结果的概率。



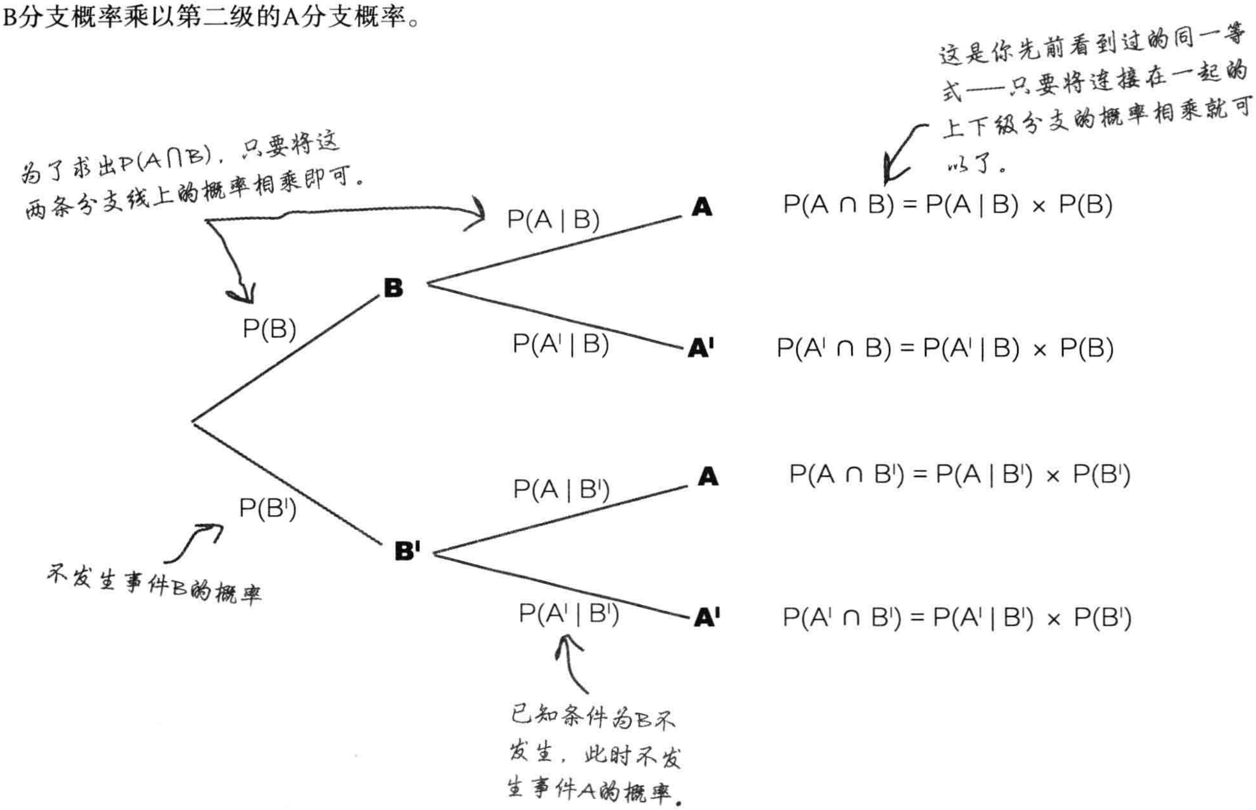
第一级分支线上标出各种结果的概率，因此“停球结果为黑”的概率为18/38，即0.474；第二级分支线上标出**已知所连接的上一级结果的情况下**的第二级结果的概率。若已知停球位置为黑色，则停球位置为奇数的概率为8/18，即0.444。

利用概率树还能计算条件概率

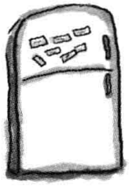
概率树不仅能帮助你以图形方式表示概率，还能帮助你计算概率。

让我们先从总体上看看概率树如何做到这一点。下面又是一幅概率树，其分支数目与前面的例子中的分支数目不一样。它显示了两级事件：A和A′以及B和B′。A′表示A中不涵盖的任何可能事件，B′表示B中不涵盖的任何可能事件。

将一个概率乘以下一级分支概率，就可以求出包含相交情况的概率。例如，假定要求P(A∩B)，可以用P(B)乘以P(A | B)，即，用第一级的B分支概率乘以第二级的A分支概率。



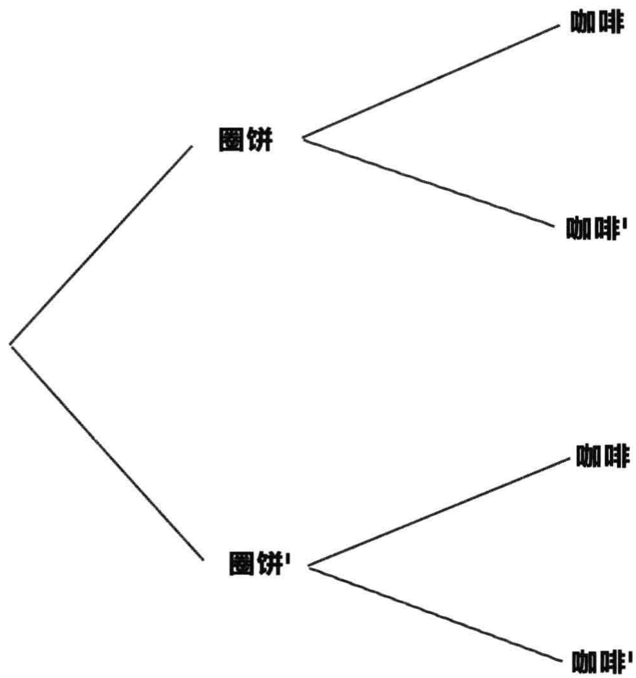
使用概率树得出的结果和以前的算法相同，用不用随便你。画概率树很费时间，但它是一种以图形体现条件概率的途径。

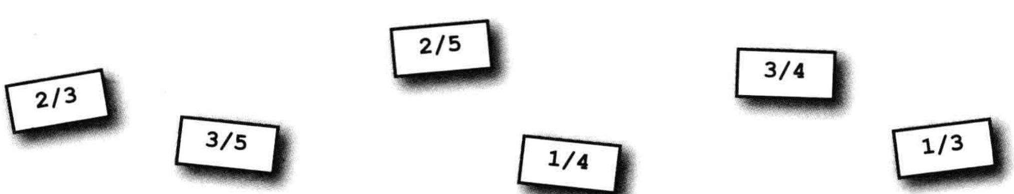


概率磁贴

邓肯圈饼店正在调查客户购买油炸圈饼和咖啡的概率。他们画了一幅概率树，用磁贴标上了各种概率。突然一阵怪风刮来，概率磁贴转眼不知所踪。你的任务就是将各个概率磁贴放回概率树。下面是一些线索。

P(圈饼) = 3/4　　P(咖啡 | 圈饼′) = 1/3　　P(圈饼 ∩ 咖啡) = 9/20





概率树使用诀窍

1.分出层级

努力分出需要计算的概率的不同层级。例如，如果给定的条件概率为P(A | B)，则可能需要在第一级中涵盖B，在第二级中涵盖A。

2.填写已知信息

如果已知部分概率，则将这些概率写入概率树上的相应位置。

3.记住：每一级分支的概率总和为1

如果将从同一个点上衍生出来的所有分支的概率加起来，总和应该等于1。记住：$P(A) = 1 - P(A')$。

4.记住公式

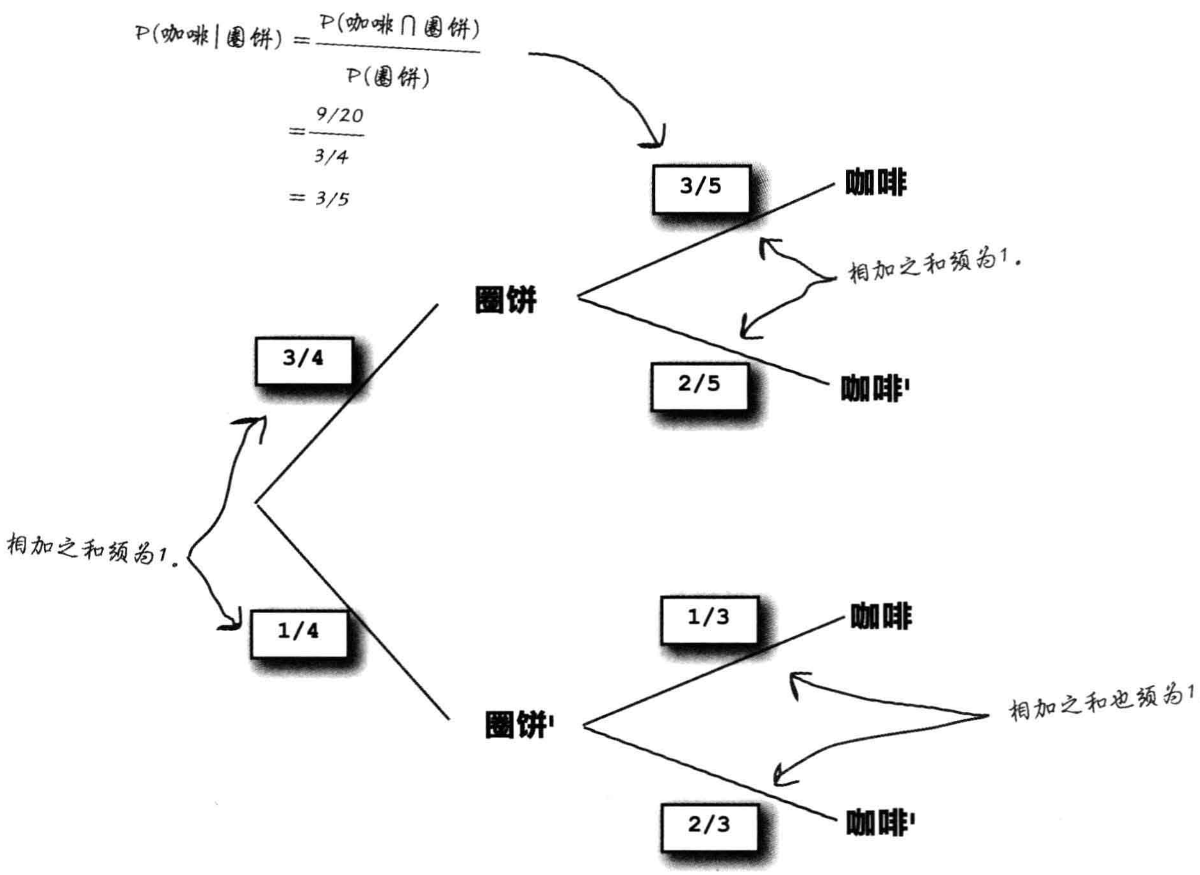
通过下列计算式可求出大多数其他概率：

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

概率磁贴解答

邓肯圈饼店正在调查客户购买油炸圈饼和咖啡的概率。他们画了一幅概率树，用磁贴标上了各种概率。突然一阵怪风刮来，概率磁贴转眼不知所踪。你的任务就是将各个概率磁贴放回概率树。下面是一些线索。

P(圈饼) = 3/4　　　P(咖啡 | 圈饼') = 1/3　　　P(圈饼 ∩ 咖啡) = 9/20





练习

邓肯圈饼店的工作还没有彻底完成！既然已经填好了概率树，请用概率树计算一些概率。

1. P(圈饼')

2. P(圈饼' ∩ 咖啡)

3. P(咖啡' | 圈饼)

4. P(咖啡) ← 提示：买咖啡的方式有几种？（你可以既买咖啡又买圈饼，也可以只买咖啡不买圈饼）

提示：也许你的某些答案能给你带来帮助。

5. P(圈饼 | 咖啡)

练习解答

你的任务是用填写完毕的概率树算出某些概率。

1. P(圈饼′)

1/4

从概率树上可以读出这个数。

我们已经知道

P(圈饼) = 3/4,

因此P(圈饼′)肯定是1/4。

2. P(圈饼′ ∩ 咖啡)

1/12

用P(圈饼′)乘以P(咖啡 | 圈饼′)可以得出这个数。我们刚才已经求出P(圈饼′) = 1/4,再从概率树上看出P(咖啡 | 圈饼′) = 1/3,
二者相乘即得1/12。

3. P(咖啡′ | 圈饼)

2/5

我们可以从概率树上读出这个数。

4. P(咖啡)

8/15

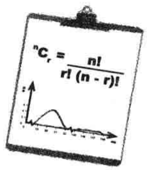
这个概率有些棘手，要是还没有算出来也不要担心。

为了求出P(咖啡)，我们需要将P(咖啡 ∩ 圈饼)和P(咖啡 ∩ 圈饼′)加起来，即：1/12 + 9/20 = 8/15。

5. P(圈饼 | 咖啡)

27/32

要求这个概率，必须先求出P(咖啡)。
P(圈饼 | 咖啡) = P(圈饼 ∩ 咖啡)/P(咖啡)，
即：(9/20) / (8 / 15) = 27/32。



重要统计量

条件

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

世上没有傻问题

问： 我仍然不清楚$P(A \cap B)$和$P(A|B)$的差别。

答： $P(A \cap B)$是A和B同时发生的概率。根据这个概率无法假设其中一个事件是否已经发生。必须在不作任何假设的情况下，求出两种事件的发生概率。

$P(A|B)$是以事件B为条件，求事件A的发生概率。也就是说，你假定事件B已经发生，然后根据这个假设算出事件A的发生概率。

问： 这么说$P(A|B)$和$P(A)$是一样的喽？

答： 不对，二者代表不同的概率。在计算$P(A|B)$的时候，必须假设事件B已经发生；而在计算$P(A)$的时候，可以不作此类假设。

问： $P(A \mid B)$和$P(B \mid A)$一样吗？看上去挺相似哦。

答： 这是个常见错误，可实际上它们是完全不一样的概率。$P(A|B)$是假定B已经发生，在此情况下A的发生概率；$P(B|A)$是假定A已经发生，在此情况下B的发生概率。二者所求的是不同已知条件下的不同事件的概率。

问： 概率树比维恩图更好用吗？

答： 两种图形都是以图形表示概率的途径，各有其妙处。维恩图的用处在于能指出基本概率及各种关系；概率树的用处则在于条件概率的计算。具体使用哪种图形取决于你要解决的问题。

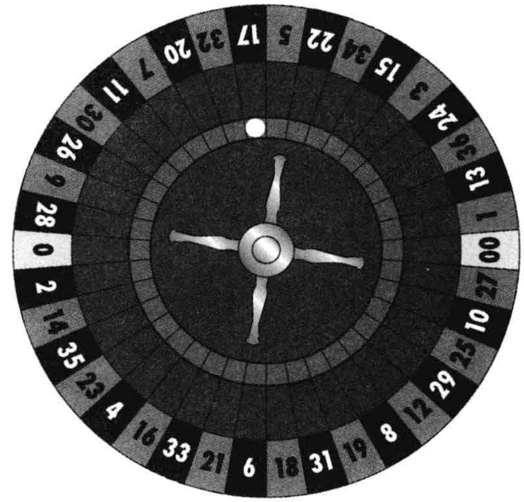
问： 概率树上的分支有层级数目限制吗？

答： 理论上没有限制。你可能会在实践中发现，超大型概率树十分难以驾驭，但尽管如此，你还是会感到驾驭超大型概率树比脱离概率树进行繁复计算来得容易。

问： 如果A与B互斥，那么$P(A \mid B)$结果如何？

答： 如果A与B互斥，则$P(A \cap B)=0$且$P(A|B)=0$。这可以理解，因为当A与B互斥时，两个事件不可能同时发生。如果我们假定事件B已经发生，则事件A不可能发生，因此$P(A|B) = 0$。

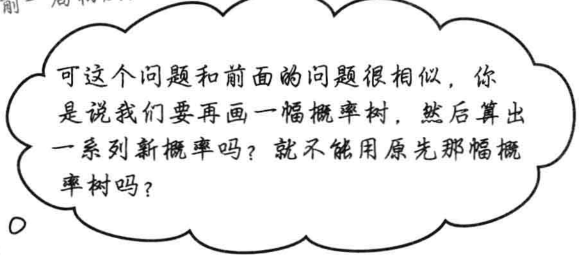
真倒霉！



在知道小球会停在黑色球位上后，你下了一注，赌小球会停在偶数球位上。真倒霉，小球停在了17上 —— 你又输掉了一些筹码。

也许我们可以再来一局，赢回一些筹码。这一次，庄家说小球会停在偶数球位上 —— 这个球位同时为黑色的概率是多少？

情况和前一局相反。

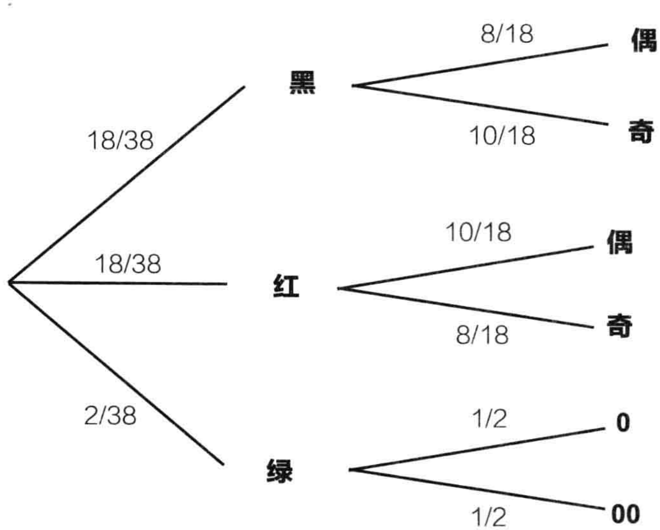






可以再次使用已经用过的算式。

上一个任务是算出P(偶|黑)，我们可以利用为了解决上一个问题而算出的概率来计算P(黑|偶)。下面是我们前面用过的概率树：



利用已有概率求P(黑|偶)

那么如何求P(黑|偶)？即使无法从概率树上直接看出这个概率，也还有办法通过已知概率算出这个概率。我们所需要做的是查看已知概率，然后设法用这些已知概率算出我们还不知道的概率。

让我们先分析要求的最终概率。

利用已有的概率，求出需要的概率。

利用求条件概率的公式，得出：

$$P(\text{黑}|\text{偶}) = \frac{P(\text{黑} \cap \text{偶})}{P(\text{偶})}$$

只要能求出P(黑∩偶)和P(偶)的概率，就能将这些概率代入公式，算出P(黑|偶)。我们需要通过一些过程求出这些概率。

觉得有困难？别担心，我们会指导你完成这个计算。

第1步：求P(黑∩偶)

让我们先算公式的第一部分：P(黑∩偶)。

动动笔

查看上一页的概率树，如何通过概率树求出P(黑∩偶)？

提示：P(黑∩偶) = P(偶∩黑)

动动笔解答

查看背面的概率树，如何利用它算出P(黑∩偶)？

将P(黑)与P(偶|黑)相乘，可求出P(黑∩偶)，即：

$$P(黑 \cap 偶) = P(黑) \times P(偶|黑)$$

$$= \frac{\cancel{18}}{38} \times \frac{10}{\cancel{18}}$$

$$= \frac{10}{38}$$

$$= \frac{5}{19}$$

我们得到了什么？

我们希望求出P(黑|偶)的概率，为此先求：

$$P(黑|偶) = \frac{P(黑 \cap 偶)}{P(偶)}$$

这两个量相等……

到现在为止，我们还只是涉及了公式的第一部分：P(黑∩偶)，而你已经了解如下算法：

$$P(黑 \cap 偶) = P(黑) \times P(偶|黑)$$

由此可得出

$$P(黑|偶) = \frac{P(黑) \times P(偶|黑)}{P(偶)}$$

于是，我们可以用P(黑)×P(偶|黑)代替原公式中的P(黑∩偶)。

下一步我们求P(偶)。

动动脑

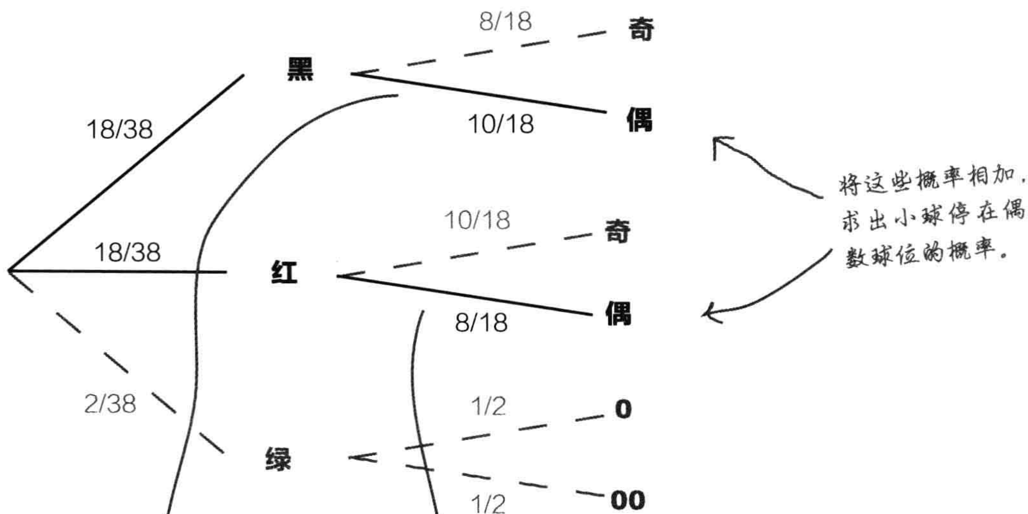
再看看166页的概率树，你觉得我们该如何利用概率树求出P(偶)？

第2步：求P(偶)

接下来求小球停在偶数球位的概率，我们可以想想发生这种结果的所有方式，据此求解。

小球停在偶数球位上的情况包括：球位既是黑色又是偶数，或者球位既是红色又是偶数。这两种情况就是小球停在偶数球位上的方式。

这表示我们可以将P(黑∩偶)与P(红∩偶)相加，得出P(偶)。也就是是说，我们将“既是黑色又是偶数的球位”的概率与“既是红色又是偶数的球位”的概率相加。概率树上的相应分支以黑色实线突出标示。



得出：

$$
\begin{aligned}
P(偶) &= P(黑\cap偶) + P(红\cap偶) \\
&= P(黑) \times P(偶|黑) + P(红) \times P(偶|红) \\
&= \frac{18}{38} \times \frac{10}{18} + \frac{18}{38} \times \frac{8}{18} \\
&= \frac{18}{38} \\
&= \frac{9}{19}
\end{aligned}
$$

小球停在偶数球位的方式

这些概率取自概率树。

步骤3：求P(黑|偶)

你还记得最初的问题吗？我们曾想求P(黑|偶)。其中：

$$P(黑|偶) = \frac{P(黑 \cap 偶)}{P(偶)}$$

一开始求的是P(黑∩偶)：

$$P(黑 \cap 偶) = P(黑) \times P(偶|黑)$$

接着求出P(偶)的表达式：

$$P(偶) = P(黑) \times P(偶|黑) + P(红) \times P(偶|红)$$

我们刚才用概率树算过这个结果。

将这些式子合并就可以利用概率树上的概率值计算P(黑|偶)：

$$\begin{aligned} P(黑|偶) &= \frac{P(黑 \cap 偶)}{P(偶)} \\ &= \frac{P(黑) \times P(偶|黑)}{P(黑) \times P(偶|黑) + P(红) \times P(偶|红)} \\ &= \frac{5}{19} \div \frac{9}{19} \\ &= \frac{5}{\cancel{19}} \times \frac{\cancel{19}}{9} \\ &= \frac{5}{9} \end{aligned}$$

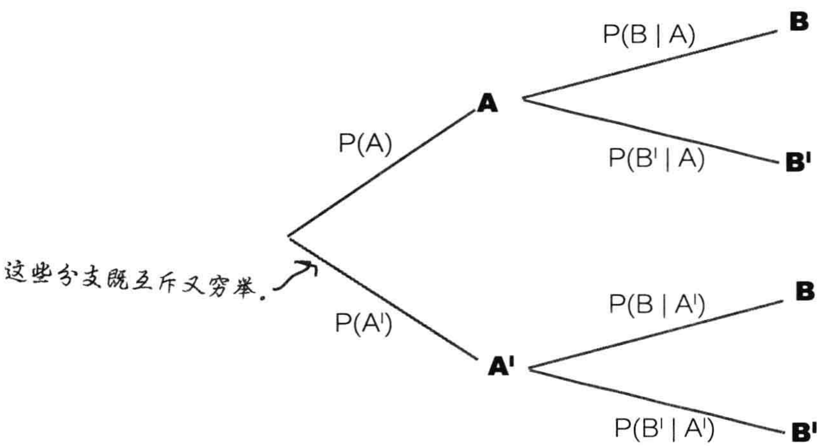
我们之前计算过，所以可以代入我们的各个结果。

这说明我们现在找到了利用已知概率求解新条件概率的方法——这就能帮助我们解决更多错综复杂的概率问题了。

让我们看看如何推而广之。

上一页的结果可以推广到其他问题

假想你有一幅概率树，上面显示了事件A和事件B的概率，假定已知每个分支的概率如下：



现在，假设你要求P(A|B)，并且知道上面的概率树上所显示的信息。请问如何使用已知概率求出P(A|B)？

我们可以从以前算过的公式开始：

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

我们必须求出这两个概率，才能得出P(A | B)。

现在，可以用概率树上的概率求出P(A∩B)，换句话说，我们可以使用下式计算P(A∩B)：

$$P(A \cap B) = P(A) \times P(B \mid A)$$

但如何求P(B)呢？

动动脑

好好观察概率树上的概率。如何利用这些概率求出P(B)?

利用全概率公式求解P(B)

让我们使用之前求解P(偶)的相同步骤求解P(B)。我们需要将想得到的事件的所有可能发生方式的概率相加。

事件B有两种发生方式：与事件A一起发生；不与事件A一起发生。即可以利用下式求出P(B)：

$$P(B) = P(A \cap B) + P(A' \cap B)$$

把这两个交集相加，得出P(B)。

我们可以根据从概率树上得知的概率，重写这个式子：

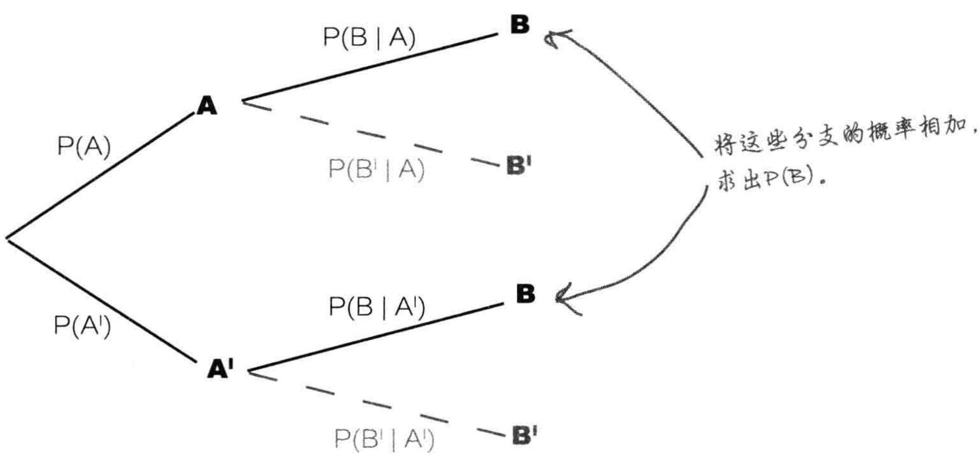
$$P(A \cap B) = P(A) \times P(B \mid A)$$

$$P(A' \cap B) = P(A') \times P(B \mid A')$$

得出：

$$\mathbf{P(B) = P(A) \times P(B \mid A) + P(A') \times P(B \mid A')}$$

这个公式有时被称为**全概率公式**，因为它提供了一种方法：根据条件概率计算一个特定事件的全概率。



既然已经求出P(A ∩ B)与 P(B)的表达式，就可以将这两个式子放在一起，得出P(A | B)的表达式。

认识贝叶斯定理

首先，我们想从概率树上已知的概率求出P(A|B)，我们已知P(A)，且已知P(B|A) 和 P(B|A′)。现在所需要的是一个求解条件概率的通用表达式，该公式是已知条件即P(A|B)的逆运算。

我们先算：

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

经过代换，这个公式……

我们在127页求出 $P(A \cap B) = P(A) \times P(B|A)$，又在前一页求出 $P(B) = P(A) \times P(B|A) + P(A') \times P(B|A')$。

将以上两个结果代入公式，得出：

$$P(A \mid B) = \frac{P(A) \times P(B \mid A)}{P(A) \times P(B \mid A) + P(A') \times P(B \mid A')}$$

……变成了这个公式。

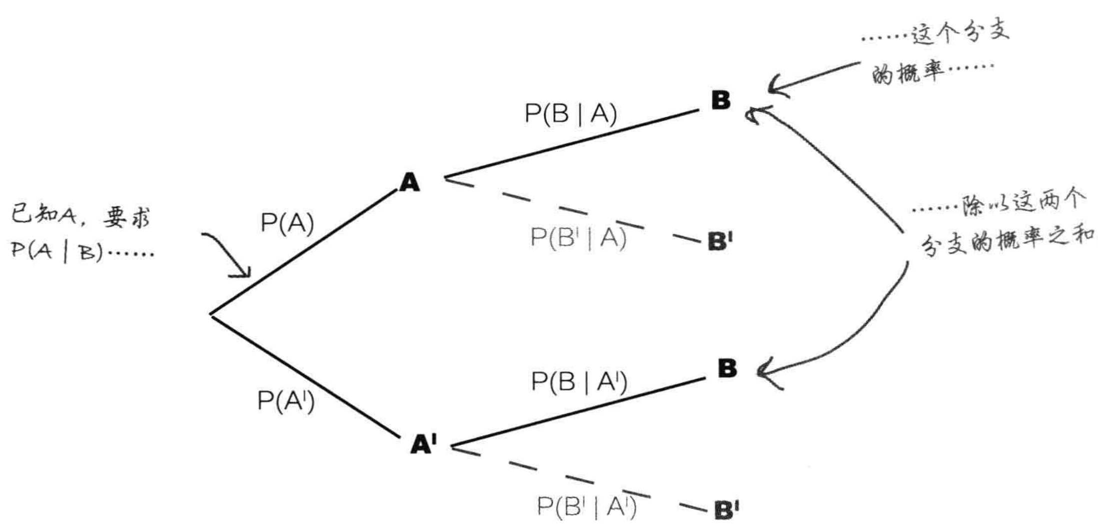
这就是所谓的**贝叶斯定理**。该定理提供了一种计算逆条件概率的方法，在你无法预知每种概率的情况下，它十分有用。



放轻松

贝叶斯定理是概率理论中最难掌握的部分之一。

若是看着觉得复杂，别担心，它计算复杂结果的能力也一样强。尽管公式棘手，我们却能借助图形得到帮助。



加强练习

芒芒游戏公司正在测试两种新游戏，他们邀请一群志愿者选择自己最喜欢玩的游戏，玩好以后告诉芒芒公司对游戏的满意程度。

80%的志愿者选择了游戏1，20%的志愿者选择了游戏2。在游戏1玩家中，有60%的人觉得好玩，40%觉得不好玩。而游戏2玩家中有70%的人觉得好玩，30%的觉得不好玩。

你的第一个任务就是填写这一例子的概率树。

芒芒公司随机挑选了一名志愿者，问她游戏是否好玩，她说好玩。这位志愿者觉得她所玩的这款游戏好玩时，她玩游戏2的概率有多大？请使用贝叶斯定理。

提示：某人选择游戏2并感到满意的概率是多大？

某人无论玩哪种游戏都感到满意的概率有多大？

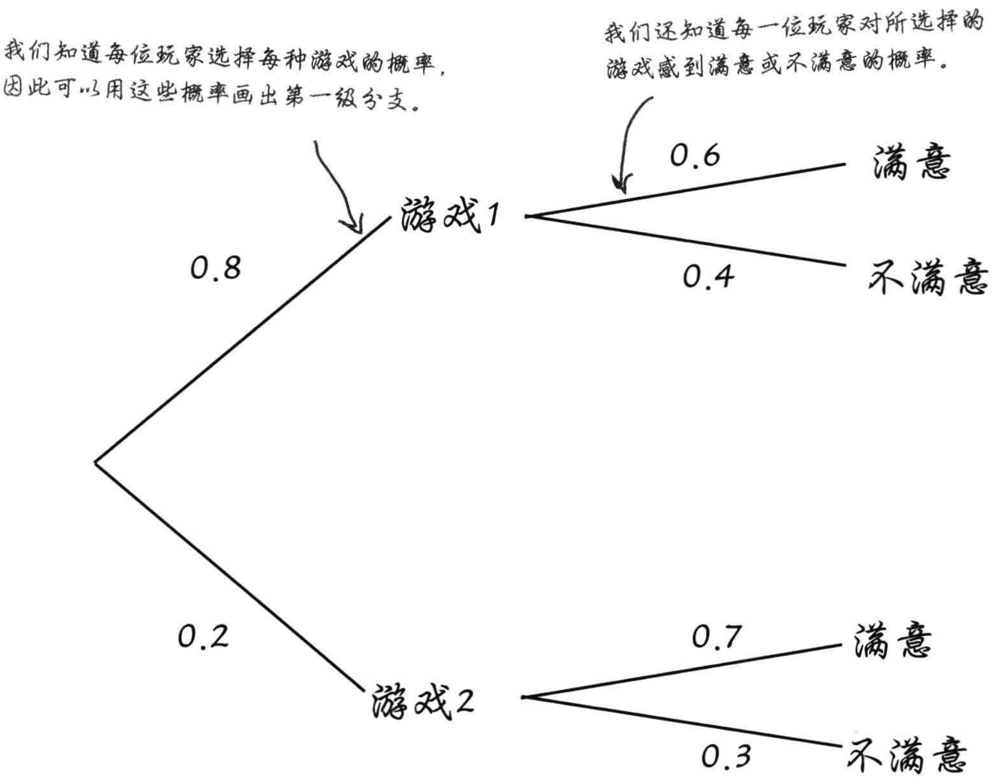
只要想通这两个问题，就能用贝叶斯定理求出正确的答案。

加强练习解答

芒芒游戏公司正在测试两种新游戏，他们邀请一群志愿者选择自己最喜欢玩的游戏，玩好以后告诉芒芒公司对游戏的满意程度。

80%的志愿者选择了游戏1，20%的志愿者选择了游戏2。在游戏1玩家中，有60%的人觉得好玩，40%觉得不好玩。而游戏2玩家中有70%的人觉得好玩，30%的觉得不好玩。

你的第一个任务就是填写这一例子的概率树。



芒芒公司随机挑选了一名志愿者，问她游戏是否好玩，她说好玩。这位志愿者觉得她所玩的这款游戏好玩时，她玩游戏2的概率有多大？请使用贝叶斯定理。

我们要用贝叶斯定理求出P(游戏2 | 满意)。公式如下：

$$P(游戏2 \mid 满意) = \frac{P(游戏2)\,P(满意 \mid 游戏2)}{P(游戏2)\,P(满意 \mid 游戏2) + P(游戏1)\,P(满意 \mid 游戏1)}$$

让我们从P(游戏2) P(满意 | 游戏2)算起

我们已经知道P(游戏2) = 0.2且P(满意 | 游戏2) = 0.7，即：

$$P(游戏2)\,P(满意 \mid 游戏2) = 0.2 \times 0.7$$

$$= 0.14$$

接下来需要求P(游戏1) P(满意 | 游戏1)。我们已经知道P(满意 | 游戏1) = 0.6以及P(游戏1) = 0.8。即：

$$P(游戏1)\,P(满意 \mid 游戏1) = 0.6 \times 0.8$$

$$= 0.48$$

将上式代入贝叶斯定理公式，得：

$$P(游戏2 \mid 满意) = \frac{P(游戏2)\,P(满意 \mid 游戏2)}{P(游戏2)\,P(满意 \mid 游戏2) + P(游戏1)\,P(满意 \mid 游戏1)}$$

$$= \frac{0.14}{0.14 + 0.48}$$

$$= \frac{0.14}{0.62}$$

$$= 0.226$$

重要统计量

全概率公式

如果有两个事件A和B，则：

$$P(B) = P(B \cap A) + P(B \cap A')$$
$$= P(A)\,P(B \mid A) + P(A')\,P(B \mid A')$$

全概率公式是贝叶斯定理的分母。

重要统计量

贝叶斯定理

如果你有n个互斥且穷举的事件：A_1至A_n，而B是另一个事件，则：

$$P(A \mid B) = \frac{P(A)\,P(B \mid A)}{P(A)\,P(B \mid A) + P(A')\,P(B \mid A')}$$

世上没有傻问题

问： 什么时候使用贝叶斯定理呢？

答： 在需要求出条件概率，且该条件概率与已知条件概率顺序相反时使用。

问： 我必须画概率树吗？

答： 你可以直接使用贝叶斯定理，也可以使用概率树进行辅助。使用贝叶斯定理更为直接快捷，但务必记住各个概率。在你忘记贝叶斯定理时，概率树很有用，不仅可以让你得出相同的结果，还能让你免于忘记每个事件所对应的概率。

问： 在轮盘赌问题中，当我们计算P(黑|偶)时，并没有将小球停在绿色球位的任何概率计算进去。我们弄错了吗？

答： 不，没有弄错。轮盘上仅有的两个绿色球位是0和00，我们并不将这两个数字计入偶数。也就是说，P(偶|绿)等于0，因此，这对计算结果没有影响。

问： 经计算，概率P(黑|偶)与P(偶|黑)相等：都是5/9。总是这样吗？

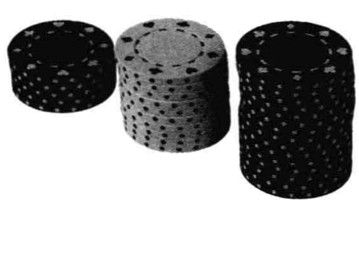
答： 的确，这里的P(黑|偶)和P(偶|黑)是一样的，但这并不表示其他情况也是如此。如果你有两个事件：A和B，不能假定P(A | B)和P(B | A)会得出相同的结果。二者指的是不同的概率，实际上，作那样的假设会让你在统计学考试中丢掉宝贵的分数。你需要使用贝叶斯定理，确保得出正确的答案。

问： 贝叶斯定理在现实生活中有用吗？

答： 实际上非常有用。例如，在计算机科学中，可以用它过滤电子邮件及检测垃圾邮件，有时它还用在医学试验中。

赢钱了！

恭喜恭喜！这次小球停在10号球位上——黑色兼偶数。你赢回了一些筹码。



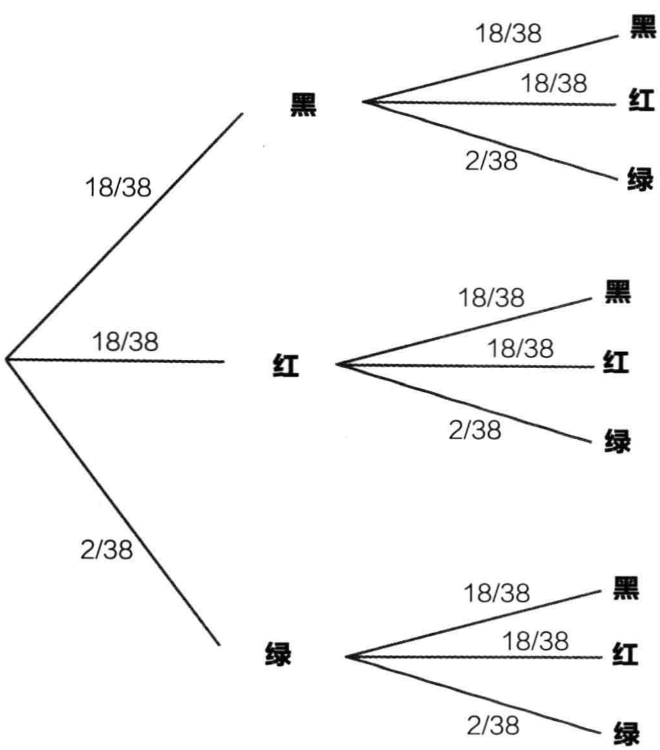


再赌最后一次

在你撤离轮盘赌之前，庄家给你的最后一注开了个大价钱：赢三倍，或赔光——如果你赌小球会连续两次停在黑色球位上，有可能赢回所有筹码。

下面是概率树。注意，“连续两次停在两个黑色球位上”的概率与166页上求解的概率有点儿不一样，在166页，我们试图计算在已知球位为黑色的条件下，停球结果为偶数球位的可能性。





如果几个事件互有影响，则为相关事件

“小球前后两次停在黑色球位上”的概率与“小球停在已知为黑色球位的偶数球位上”的概率略有区别。请看下面的概率算式：

P(偶|黑) = 10/18 = 0.556

对于P(偶|黑)来说，“停在偶数球位”的概率受到“停在黑色球位”的概率的影响，我们知道小球已经停在黑色球位上，于是利用这一点计算概率：我们查看在所有黑色球位中，有几个球位是偶数。

如果我们不知道小球已经停在黑色球位上，则概率会不一样。为了计算P(偶)：我们查看在所有的球位中，有几个球位是偶数。

P(偶) = 18/38 = 0.474

这两个概率是不一样的

P(偶|黑)得出了与P(偶)不一样的结果，换句话说，我们所得知的“球位为黑色”的信息使概率发生了改变。我们说这两个事件是**相关事件**。

如果用通用术语表达就是：如果P(A | B) 与P(A)不等，则我们说事件A与事件B是相关事件 —— 这等于说事件A与事件B的概率相互影响。



动动脑

再看一看前一页的概率树。你注意到每一级分支的特点了吗？“小球在第一局中停在黑色球位上”和“小球在第二局中停在黑色球位上”是相关事件吗？为什么？

如果几个事件互不影响，则为独立事件

并非所有事件都是相关事件，有时候，几个事件相互之间完全没有影响，无论其他事件发生与否，某个事件的发生概率总是保持不变。例如，请看P(黑)和P(黑|黑)的概率，你注意到什么了？

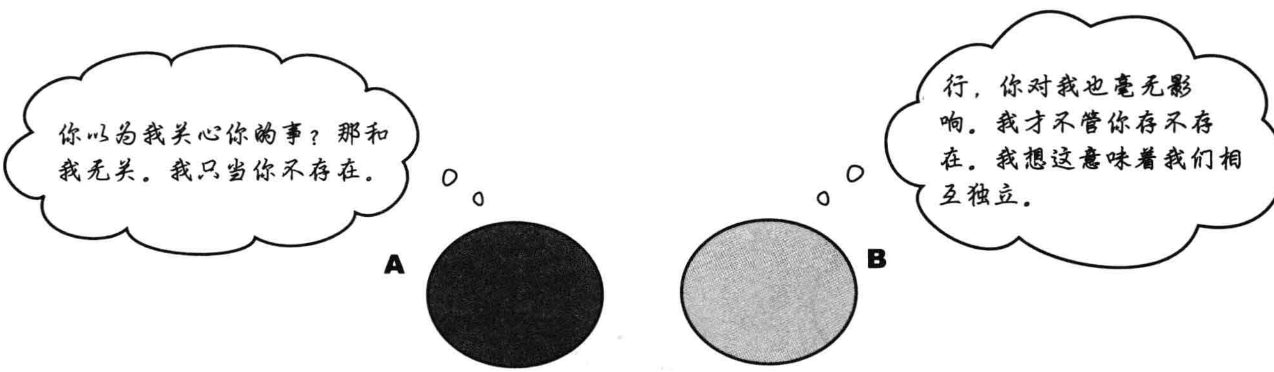
P(黑) = 18/38 = 0.474

P(黑 | 黑) = 18/38 = 0.474

这些概率相等，事件与事件相互独立。

以上两个概率数值相同，换句话说，“小球在这一局停在黑色球位上”事件对“小球在下一局停在黑色球位上”事件没有影响，这两个事件是**独立的**。

独立事件彼此之间互不影响 —— 不以任何形式相互影响对方的概率。若一个事件发生，其他事件的概率保持原样，纹丝不变。



如果事件A和事件B相互独立，则事件A的概率不受事件B的影响，换句话说，对于独立事件来说：

$$P(A \mid B) = P(A)$$

我们还能用以上公式进行独立性检验。如果你有两个事件A和B，且P(A | B) = P(A)，则事件A和事件B必然相互独立。

再谈独立事件概率计算

独立事件的其他概率也很容易计算，例如P(A ∩ B)。

我们已经知道

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

如果A和B是独立事件，则P(A | B) 与P(A)相同。即对于独立事件来说：

$$P(A) = \frac{P(A \cap B)}{P(B)}$$

或

$$\mathbf{P(A \cap B) = P(A) \times P(B)}$$

换句话说，如果两个事件相互独立，则通过将两个事件各自的概率相乘，可以算出同时发生这两件事的概率。



如果A、B是互斥事件，则二者不会是独立事件；如果A、B是独立事件，则二者不会是互斥事件。

如果A和B是互斥事件，即如果事件A发生，则事件B不发生。这意味着，A的结果会影响B的结果，于是这二者相关。

与此相似，如果A和B是独立事件，则二者不会互斥。

动动笔

现在该计算另一个概率了：“小球连续两次停在黑色球位上”的概率是多少？

动动笔解答

现在该计算另一个概率了：“小球连续两次停在黑色球位上”的概率是多少？

我们需要求P(第一局黑色 ∩ 第二局黑色)。由于这两个事件相互独立，因此：

$$18/38 \times 18/38 = 324/1444$$

$$= 0.224 \text{（保留三位小数）}$$

世上没有傻问题

问： 独立事件和互斥事件有何差别？

答： 假想你有两个事件：A和B。如果A和B互斥，则在事件A发生时，B无法发生。同样，如果事件B发生，则A无法发生。换句话说，二者不可能同时发生。

如果A和B是独立事件，则A的结果对B的结果没有影响，同时B的结果对A的结果没有影响。二者各自的结果对对方没有影响。

问： 两个事件必须同时为独立事件吗？能不能其中一个事件是独立事件，而另一个事件是相关事件？

答： 不能，两个独立事件指的是“相互”独立，因此不可能一个是相关事件，另一个是独立事件。

问： 轮盘赌中的每一局都是独立事件吗？为什么？

答： 没错，都是独立事件。轮盘的每一次转动都不会前后影响。小球在每一局中停在红色、黑色或绿色球位上的概率是不变的。

问： 你已经演示过如何使用概率树论证独立事件。如何使用维恩图判断几个事件是否相互独立？

答： 维恩图的确不是体现相关性的最好方法。维恩图在检验交集、表现互斥事件方面表现极佳，但在表现独立性方面效果并不好。

重要统计量

独立性

如果A和B相独立，则：

$$P(A \mid B) = P(A)$$

如果上式对任何两个事件成立，则这两个事件必为独立事件。同时：

$$P(A \cap B) = P(A) \times P(B)$$



瑜伽班与游泳班案例

Head First健身俱乐部为自己能为每一位前来健身的人找到合适的班级感到自豪，这正是俱乐部风靡老中少健身者的原因。

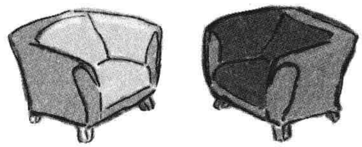
健身俱乐部目前正在动脑筋，为的是最有效地推销它新开设的瑜伽班，他们想知道，是否参加游泳班的人更有可能参加瑜伽班。“也许我们可以给游泳班学员一些折扣，鼓励他们参加瑜伽班。”

首席执行官不同意。“我想你们错了”，他说，“我想参加游泳班的人和参加瑜伽班的人是相互独立的，我不认为参加游泳班的人比其他人更有可能参加瑜伽班。”

他们调查了96个人，问他们是否参加游泳班或瑜伽班。在这96个人中，有32人参加瑜伽班，72人参加游泳班。有24人最为积极，两个班都参加了。

那么，谁对谁错？瑜伽班和游泳班是相关，还是相互独立？

面对面



今夜谈：相关与独立探讨相互间的差异

相关： 独立老兄，很高兴看到你露面。我早就想逮住你问问了。

独立： 是吗，相关老兄？为什么呢？

相关： 哦，我听说你总给菜鸟统计师惹麻烦，没有你的时候，他们干得很顺利，可是只要你一来，天啦，错误概率就满天飞啊！小∩尤其对你有意见。

独立： 我有点儿伤心呢，小∩居然说我的坏话，我以为自己让他过得轻松了呢。他想算出发生两个独立事件的概率？容易！只要把两个事件的概率相乘，就大功告成了。

相关： 就是你这种简单的态度给人们带来了麻烦。他们想：“嘿，这位独立老兄看起来挺简单，我就用他来算这个概率。”然后呢，你知道的，∩把所有的概率胡乱混在一起。这可不是处理相关事件的正确方法。

独立： 你言过其实了。即使人们决心用我而不用你，也不见得会引起多大差别。

相关： 你不明白事情的严重性。如果人们按照你的方式计算∩概率，而事件是相关事件，那么他们肯定会得出错误答案，这可不太好。对于相关事件，只有在考虑小|的时候——小|代表已知条件，你才能得出正确答案。

独立： 我不能说自己给了他们很多关注。对于独立事件来说，概率结果都是一样的。

相关：

你又来了——你把事情看得过于简单。好吧，我已经说得够多了。我想人们应该首先想到我，而不是你，才能把所有这些问题都搞清楚。

独立：

是吗？怎么会这样呢？

相关：

彻底想清楚事情是不是相关事件。我来举个例子：假设你有一副牌，共52张，其中13张是方块。想象你随机抽了一张牌，发现是方块。发生这个事件的概率有多大？

独立：

这简单。13/52，或者说1/4。

相关：

再抽第二张牌会怎样？抽出第二张方块的概率是多大？

独立：

一样嘛，对不？1/4。

相关：

不对！这些事件是相关的。你不能再认为这副牌里有13张方块——你已经抽掉了一张，因此只剩下51张牌，其中方块12张。概率变为12/51，或者说4/17。

独立：

不公平，我以为你把第一张牌放回去了！

那就意味着抽出方块的概率和以前一样，我就是对的。这些事件应该是独立的。

相关：

但它们不是。当人们首先想到你的时候，他们就会作出许多不恰当的假设。这就难怪小∩乱成一团了。

独立：

哦，谢谢你给我讲这些，相关老兄，很高兴我们有机会把事情讲清楚。

相关：

别放在心上，下次考虑事情小心全面一些就行了。

破解：瑜伽班与游泳班案例

瑜伽班和游泳班是相关的还是独立的？



首席执行官是对的——两个班是独立的。

下面是他了解的信息：

96人中有32人上瑜伽班，因此：

$$P(瑜伽) = 1/3$$

72人上游泳班，因此：

$$P(游泳) = 3/4$$

24人两个班都上，因此：

$$P(瑜伽 \cap 游泳) = 1/4$$

可我们怎么知道这两个班是相互独立的呢？让我们将P(瑜伽)和P(游泳)相乘，看看结果。

$$\begin{aligned} P(瑜伽) \times P(游泳) &= 1/3 \times 3/4 \\ &= 1/4 \end{aligned}$$

由于这个结果等于P(瑜伽∩游泳)，于是我们知道两个班级是相互独立的。

相关还是独立？

下面是一些情况和事件，请说出哪些是相关事件，哪些是独立事件。

	相关	独立
掷出硬币，连续两次正面朝上。	☐	☐
从抽屉里拿袜子，直到找出一双。	☐	☐
从一盒巧克力中随机拿巧克力，连续两次拿到黑巧克力。	☐	☐
从一副牌里拿出一张牌，然后抽出另一张牌。	☐	☐
从一副牌里抽出一张牌，将这张牌放回去，然后抽出另一张牌。	☐	☐
在星期二（已知条件）下雨。	☐	☐

相关还是独立？解答

下面是一些情况和事件，请说出哪些是相关事件，哪些是独立事件。

掷第二枚硬币的概率不受掷第一枚硬币的影响。

在取出一只袜子后，下一次取袜子时，原来的袜子数就减少了，这会影响概率。

不会由于是星期二而更有可能下雨或不下雨，因此二者是独立事件。

	相关	独立
掷出硬币，连续两次正面朝上。	☐	☑
从抽屉里拿袜子，直到找出一双。	☑	☐
从一盒巧克力中随机拿巧克力，连续两次拿到黑巧克力。	☑	☐
从一副牌里拿出一张牌，然后抽出另一张牌。	☑	☐
从一副牌里抽出一张牌，将这张牌放回去，然后抽出另一张牌。	☐	☑
在星期二（已知条件）下雨。	☐	☑

赢钱了！赢钱了！

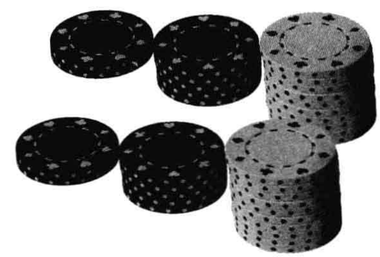


轮盘连转两次，小球都落在30号红色球位上 —— 你赢了双倍。

你已经在肥蛋赌场的轮盘赌桌上学了大量概率知识，这些知识将在赌场中的其他赌博游戏中派上用场。不过，可惜哦，你收入荷包的筹码可不够多哦。

[肥蛋赌场消息：
我们松了一口气。]

真是太好了，我们知道赢取各种赌法的几率。不过，除了概率，是不是该多懂一些，才能智胜赌场？





除了赢钱概率，还需要知道赢钱的金额，以便决定是否该冒险下注。

对于一个概率极低的事件，如果回报足以弥补所承担的风险，则值得押上一注。在下一章中，我们将看看如何将回报纳入概率计算式，帮助我们作出更有根据的赌博决策。

健忘的聚餐者

三位健忘的朋友决定外出用餐，但他们忘了打算在哪儿会面了。弗莱德决定掷硬币帮忙：如果正面着地，则去蒂勒餐厅；如果反面着地，则去意大利餐厅。乔治也掷了硬币：正面着地，去意大利餐厅；反面着地，去蒂勒餐厅。罗恩决心只去意大利餐厅，因为他喜欢那家餐厅的食物。

三位朋友见面的概率有多大？其中一位单独用餐的概率有多大？



练习

下面再增加一些轮盘赌概率，供你练习。

1. 已知停球位置为黑色，求小球停在数字17的概率。

2. 小球连续两次停在22球位的概率。

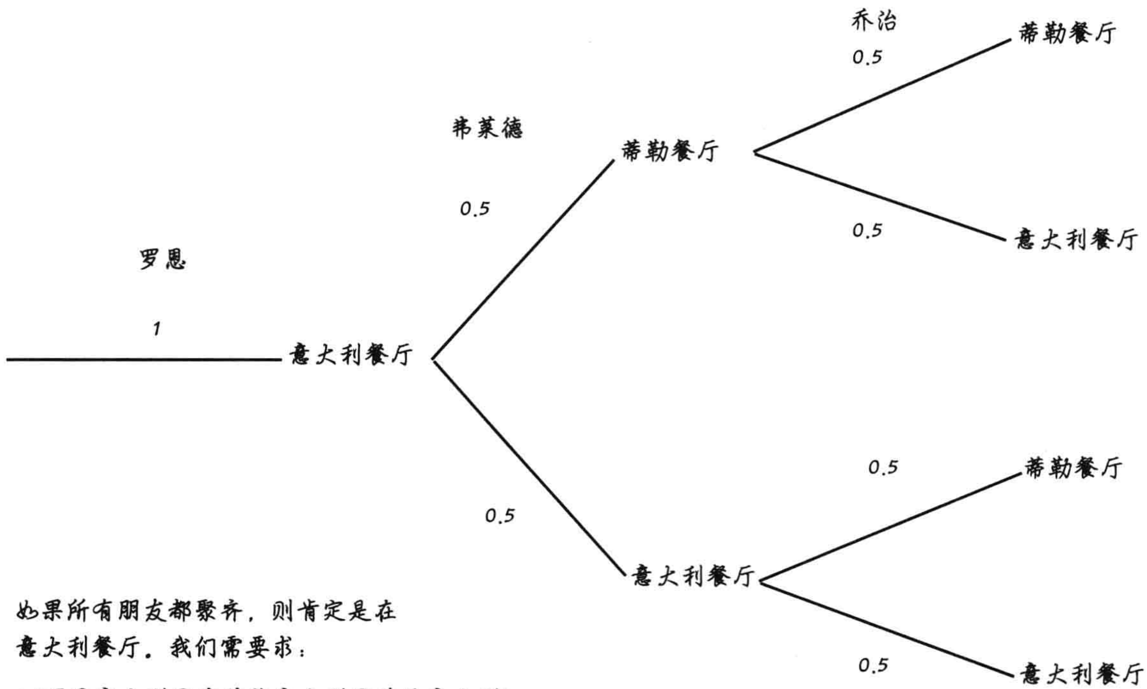
3. 已知停球位置为红色，求小球停在编号大于4的球位的概率。

4. 小球停在1、2、3或4的概率。

健忘的聚餐者

三位健忘的朋友决定外出用餐，但他们忘了打算在哪儿会面了。弗莱德决定掷硬币帮忙：如果正面着地，则去蒂勒餐厅；如果反面着地，则去意大利餐厅。乔治也掷了硬币：正面着地，去意大利餐厅；反面着地，去蒂勒餐厅。罗恩决心只去意大利餐厅，因为他喜欢那家餐厅的食物。

三位朋友见面的概率有多大？其中一位单独用餐的概率有多大？



如果所有朋友都聚齐，则肯定是在意大利餐厅。我们需要求：

P(罗恩意大利 ∩ 弗莱德意大利 ∩ 乔治意大利)

$= 1 \times 0.5 \times 0.5 = 0.25$

有1个人单独用餐的情况是：弗莱德和乔治去蒂勒餐厅；弗莱德去蒂勒餐厅，而乔治去意大利餐厅；或乔治去蒂勒餐厅，而弗莱德去意大利餐厅。

$(0.5 \times 0.5) + (0.5 \times 0.5) + (0.5 \times 0.5) = 0.75$

练习解答

下面再增加一些轮盘赌概率，供你练习。

1. 已知停球位置为黑色，求小球停在数字17的概率。

黑色球位有18个，其中之一编号17。

$P(17 \mid 黑) = 1/18 = 0.0556$ (保留三位小数)

2. 小球连续两次停在22球位的概率。

我们需要求$P(22 \cap 22)$，由于这些事件是独立事件，因此这个式子等于$P(22) \times P(22)$。停球结果为22的概率是1/38,因此：

$P(22 \cap 22) = 1/38 \times 1/38 = 1/1444 = 0.00069$ (保留五位小数)

3. 已知停球位置为红色，求小球停在编号大于4的球位的概率。

$P(大于4 \mid 红) = 1 - P(4或4以下 \mid 红)$

小于4的红色球位有2个，因此：

$1 - (1/18 + 1/18) = 8/9 = 0.889$ (保留三位小数)

4. 小球停在1、2、3或4的概率。

每个球位的概率为1/38，因此所述事件的概率为

$4 \times 1/38 = 4/38 = 0.105$ (保留三位小数)

5 离散概率分布的运用

善用期望

好了，从树上摔下来不是我们期望的结果，不过，对这种事我们必须看远一些。



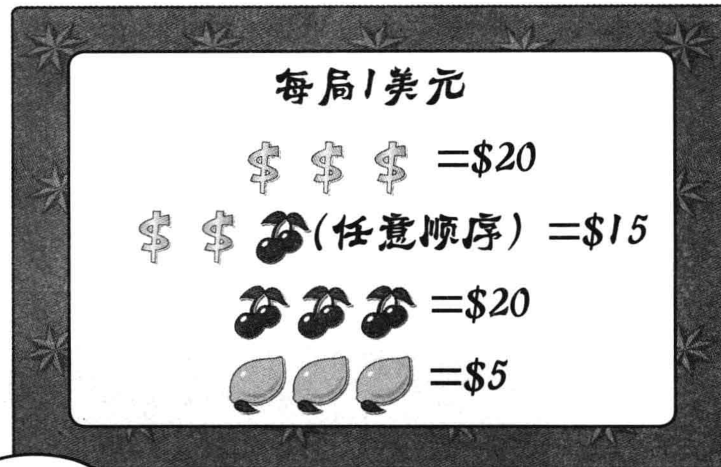
意外从天而降，未来如何演变？

前文讲到如何通过概率得知发生某些事件的可能性的大小。可惜概率并非万能，它无法指出所发生的这些事情的**整体影响**，也无法指出这种整体影响对你的具体影响。不错，你有时会在轮盘赌中大赚特赚，但你赚到的钱真的填得平那些赔掉的钱吗？在本章中，我们将讲述如何利用概率**预测长期结果**，以及如何量度这些预测结果的**确定性**。

重回肥蛋赌场

你曾经痴迷于老虎机忽闪忽闪的灯光吗？好吧，你走运了，肥蛋赌场有一长排灯光闪闪的老虎机等着你来玩呢，让我们来到其中一台老虎机前，以1美元一局（拉一次杆）的赌本玩起来。没准儿你会大发一笔！

这台老虎机有三个窗口，如果三个窗口全部恰到好处地亮起来，成堆的硬币就会滚滚而下。



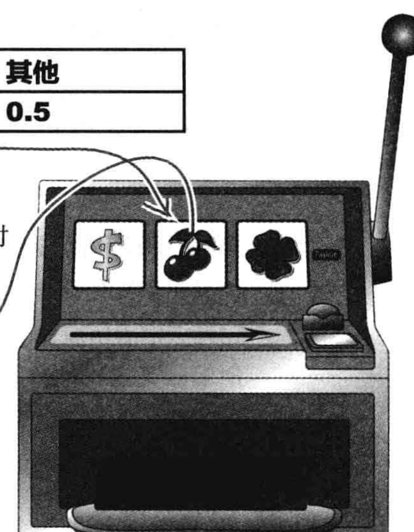


大把赢钱当然充满诱惑，但在开赌之前，我希望先搞清楚撞上这些组合的概率。

似乎我们是有办法算一算的。下面是一个特定图像出现在一个特定窗口中的概率。

$	樱桃	柠檬	其他
0.1	0.2	0.2	0.5

这三个窗口相互独立，即每个窗口中出现的图像对其他窗口中出现的图像没有影响。



樱桃出现在这个窗口中的概率是0.2。

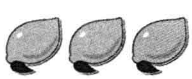
化身赌徒



看一看上一页的老虎机海报，你的任务是化身赌徒，算出海报上的各种组合的发生概率。一无所获的概率是多少？

的概率

的概率（任意顺序）

的概率

的概率

一无所获的概率

化身赌徒解答



看一看上一页的老虎机海报，你的任务是化身赌徒，算出海报上的各种组合的发生概率。一无所获的概率是多少？

$\$\ \$\ \$$ 的概率

$P(\$, \$, \$) = P(\$) \times P(\$) \times P(\$)$

$= 0.1 \times 0.1 \times 0.1$

$= 0.001$

一个窗口中出现一个美元符号的概率是0.1。

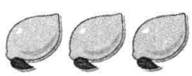
$\$\ \$$  的概率（任意顺序）

出现这种组合的情况有三种：

$P(\$, \$, \text{樱桃}) + P(\$, \text{樱桃}, \$) + P(\text{樱桃}, \$, \$)$

$= (0.1^2 \times 0.2) + (0.1^2 \times 0.2) + (0.1^2 \times 0.2)$

$= 0.006$

 的概率

$P(\text{柠檬}, \text{柠檬}, \text{柠檬}) = P(\text{柠檬}) \times P(\text{柠檬}) \times P(\text{柠檬})$

$= 0.2 \times 0.2 \times 0.2$

$= 0.008$

一个窗口中出现一个柠檬与其他两个窗口中出现柠檬是相互独立的事件，因此将这三个概率相乘。

 的概率

$P(\text{樱桃}, \text{樱桃}, \text{樱桃}) = P(\text{樱桃}) \times P(\text{樱桃}) \times P(\text{樱桃})$

$= 0.2 \times 0.2 \times 0.2$

$= 0.008$

一无所获的概率

即没有撞上任何赢钱组合的概率。

与其算出所有可能出现的赔钱方式，还不如求出P(赔钱) = 1 - P(赢钱)。

$P(\text{赔钱}) = 1 - P(\$, \$, \$) - P(\$, \$, \text{樱桃（任意顺序）}) - P(\text{樱桃}, \text{樱桃}, \text{樱桃}) - P(\text{柠檬}, \text{柠檬}, \text{柠檬})$

$= 1 - 0.001 - 0.006 - 0.008 - 0.008$

$= 0.977$

这是前面算出的四个概率。

我们可以写出老虎机概率分布

下面是老虎机的各种赢钱组合的概率。

这不过是对前面算出的概率进行汇总而已。

组合	无	柠檬	樱桃	美元/樱桃	美元
概率	0.977	0.008	0.008	0.006	0.001

这张表看上去很有用，不过我在想，我们是不是能够再深入一些？我们已经求出了每种赢钱组合的概率，但我们真正感兴趣的是能赚多少钱或者会赔多少钱。



我们不仅想知道赢钱的概率，还想知道赚钱数额——收益

目前我们是基于符号组合来写概率，这就很难一眼看出我们能赚多少，好在我们并不一定要这样写。

现在让我们放弃基于老虎机图形写概率的做法，代之以基于每一局的收益或赔付写概率。为此还需做这样一个计算：用每一个组合对应的赢金（即海报上注明的金额）减去玩一局的本金（1美元）。

组合	无	柠檬	樱桃	美元/樱桃	美元
收益	-$1	$4	$9	$14	$19
概率	0.977	0.008	0.008	0.006	0.001

若不能撞上赢钱组合，就得赔掉1美元。

同样的概率，但基于收益写出。

撞上某种赢钱组合后的收益 = 赢金 - 1美元本金。

表格给出了赢局的**概率分布**——即老虎机每一种可能收益(或赔付)所对应的概率的集合。

概率分布细细看

在推算老虎机概率时，你计算了每个赢局（或赔局）的概率，即，你计算了一个随机变量的概率分布。**随机变量**是一个可以等于一系列数值的变量，而这一系列数值中的每一个值都与一个特定概率相关联。在肥蛋赌场老虎机这个例子中，随机变量代表我们将在每一局赌局中赢得的收益。

随机变量通常用大写字母表示，如X或Y；变量能够采用的特定数值则用小写字母表示，如x或y。于是，P(X = x)则表示“变量X取特定数值x的概率”。

以下是用上述表示法表示的老虎机的概率分布：

这里的x是19。

每个组合的收益以x表示。

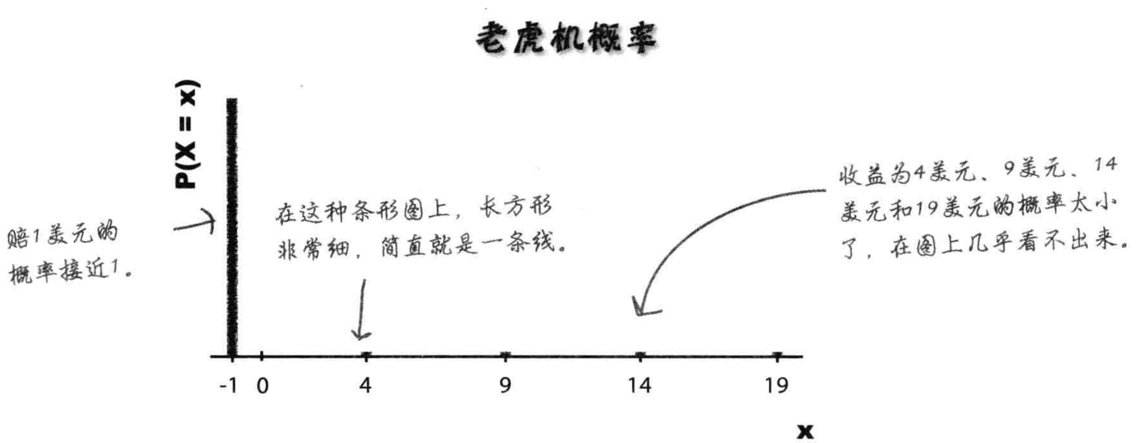
组合	无	柠檬	樱桃	美元/樱桃	美元
x	-1	4	9	14	19
P(X = x)	0.977	0.008	0.008	0.006	0.001

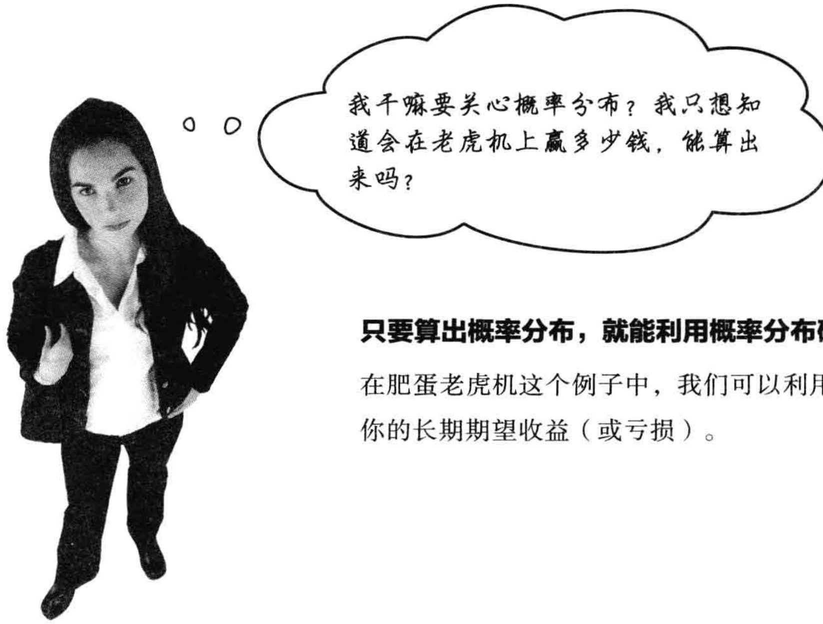
x是变量。

变量X等于9（即收益为9美元）的概率。

这里的变量具有**离散性**，即该变量只能取确定数值。

除了拟定概率分布表，我们还可以用图形来表示概率分布。下面是一张条形图，用于表示老虎机的概率。





只要算出概率分布，就能利用概率分布确定预期结果。

在肥蛋老虎机这个例子中，我们可以利用概率分布确定你的长期期望收益（或亏损）。

世上没有傻问题

问： 我们为什么不能用符号，反而要用数字呢？我可没把握是不是真的会赢那么多钱。

答： 我们可以用符号，但用数字代替符号能做更多事，因为数字可以参加计算。例如，你即将看到如何利用这些数字计算我们能够期望在每一局赌局中赢多少钱。如果只用符号的话，可作不了这样的预测。

问： 如果我想用维恩图体现概率分布，能办到吗？

答： 用这个方法体现概率分布不是特别合适。维恩图和概率树在计算概率时很有用，但对于概率分布来说，所有概率都早已计算好了。

问： 我能用任意字母表示某个变量吗？

答： 可以，只是别用乱了。最常见的情况是用字母表末尾的几个字母来表示，例如X和Y。

问： 我应该用相同的字母表示变量和数值吗？或许我该用X代表变量，y代表数值？

答： 从理论上讲这并非不可，不过在实际应用中，你会发现用不同的字母更容易引起混淆，最好坚持用相同的字母分别表示变量和数值。

问： 你说过，离散随机变量就是能精确指出其数值的变量，我倒觉得每个变量都有这种特点，难道不是吗？

答： 并非如此。在老虎机例子中，你确切地知道每一种符号组合的相应收益——确切得不能再确切，无论玩多少次，对于每一局赌局来说，可能的赢钱数值都保持不变。

但还有一些时候，你得到的是一个数值范围，这个数值范围内的任何数值都有可能出现。例如，假定要求你测量一些长度在10英寸到11英寸范围内的丝线的具体长度，那么，丝线长度完全可以是这个范围内的任何数值。

暂时不用过于担心其中区别，我们将在本书后续部分详加说明。目前，我们所研究的随机变量都将是离散性的。

期望指示预测结果……

你已经有了老虎机的收益概率分布，但现在需要知道自己能够期望获得的长期收益。为了算出这个期望数额，可以先算出在典型情况下可以期望每一局赢多少或赔多少，即可以求出统计学上的所谓**期望**。

变量X的期望和均值有点儿像，甚至连计算方法也相似，但它描述的是概率分布。为了求出期望，可将每个数值x乘以该数值的发生概率，然后将所有乘积求和。

变量X的期望通常写作E(X)，但有时候也会写作μ，也就是均值的符号。我们这样打比方吧：期望和均值是一对双胞胎，但一出生就由不同人家领养了。



$$E(X) = \mu$$

下面是E(X)的计算式：

将每个数值与其概率相乘

下面是E(X)的计算式：

$$E(X) = \sum xP(X = x)$$

得出乘积后，将所有乘积相加。

让我们用这个算式计算老虎机的收益期望。下面是所用概率分布的提示数字：

x	-1	4	9	14	19
P(X = x)	0.977	0.008	0.008	0.006	0.001

$$E(X) = (-1 \times 0.977) + (4 \times 0.008) + (9 \times 0.008) + (14 \times 0.006) + (19 \times 0.001)$$

$$= -0.977 + 0.032 + 0.072 + 0.084 + 0.019$$

$$= -0.77$$

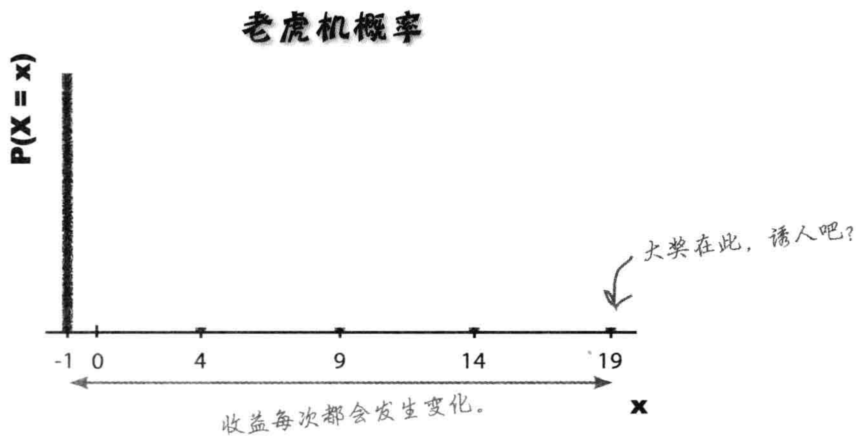
让我们用这个算式计算老虎机的收益期望。下面是所用概率分布的提示数字：

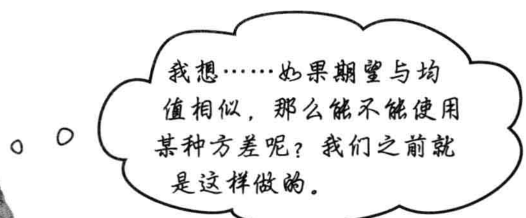
换句话说，在多次拉杆之后，你能够期望每一局赔掉0.77美元，也就是说，如果玩100次老虎机，你能够期望赔掉77美元。

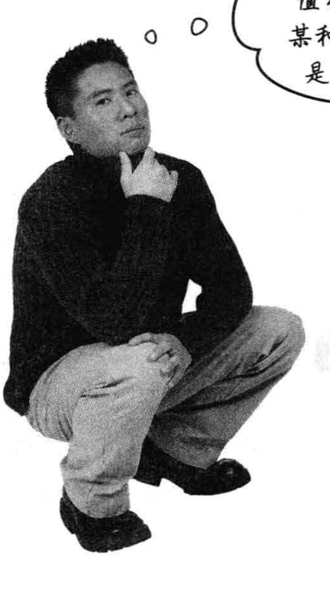
方差指示结果的分散性

期望指出每一局赌局能够期望得到的平均收益，如果每一次都赔这么多钱，那么赌博有何乐趣？谁又愿意赌博？

有理由期望每一局赌博都赔钱并不表示连一丁点儿赢大钱的希望都没有。和均值一样，期望并没有全面体现出每一局赌局有可能存在的收益变化。你觉得该怎么量度这种变化？







概率分布确实有其方差。

期望指出一个变量的典型值或平均值，但并不提供有关数值分散性的任何信息。在老虎机赌博中，如能得到分散性信息，我们将能更多地了解潜在收益的变化情况。

像第3章中的做法一样，我们可以使用**方差**来量度这种分散性。让我们看看具体做法。

方差和概率分布

先回顾一下第3章：我们计算了一批数字的方差——我们算出每个数字的 $(X-\mu)^2$，然后取所有计算结果的平均值。

类似地，我们可以算出变量X的方差，但我们不求 $(X-\mu)^2$ 的**平均值**，而是求 $(X-\mu)^2$ 的**期望**。计算公式如下：

μ是E(X)的另一种记法。

$$Var(X) = E(X-\mu)^2$$

这是方差——Var(X)是X的方差的简便记法。

我们需要求 $(X-\mu)^2$ 的期望——用哪种方法呢？

只有一个问题：如何求出 $(x-\mu)^2$ 的期望？

如何计算 $E(X-\mu)^2$？

求 $E(X-\mu)^2$ 的方法与求E(X)的方法非常相似。

计算E(X)时：取概率分布中的每一个数值，乘以其概率，然后将各个乘积相加。也就是使用下式进行计算：

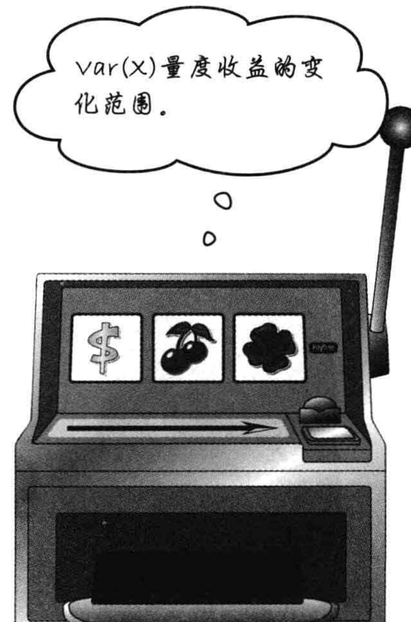
$$E(X) = \Sigma xP(X = x)$$

计算X的方差时：计算每个数值x的 $(x-\mu)^2$，用所得结果乘以相应数值x的发生概率，然后将各个结果相加。

取每一个数值x，算出 $(x-\mu)^2$；用所得结果乘以相应x的发生概率……

$$E(X-\mu)^2 = \sum(x-\mu)^2P(X = x)$$

……然后将所有乘积相加。



也就是说，你不是用x乘以其相应概率，而是用 $(x-\mu)^2$ 乘以相应x的发生概率。

让我们算算老虎机的方差

让我们看看能否用上述方法计算老虎机的方差，为此，我们用每一个值减去μ，取差的平方，然后乘以概率。提示一下，E(X)或μ等于-0.77。

老虎机概率提示

x	-1	4	9	14	19
P(X = x)	0.977	0.008	0.008	0.006	0.001

我们在204页求得 E(X) = -0.77。

$$Var(X) = E(X - \mu)^2$$

$$= (-1+0.77)^2 \times 0.977 + (4+0.77)^2 \times 0.008 + (9+0.77)^2 \times 0.008 + (14+0.77)^2 \times 0.006 + (19+0.77)^2 \times 0.001$$

$$= (-0.23)^2 \times 0.977 + 4.77^2 \times 0.008 + 9.77^2 \times 0.008 + 14.77^2 \times 0.006 + 19.77^2 \times 0.001$$

$(X-\mu)^2 \times P(X=x)$

$$= 0.0516833 + 0.1820232 + 0.7636232 + 1.3089174 + 0.3908529$$

$$= 2.6971$$

这就是说，当收益期望为-0.77时，方差为2.6971。



那标准差呢？我们也能计算吗？

就像可以算出方差一样，也可以算出概率分布的标准差。

概率分布的标准差与数据集的标准差作用相似，是一种量度数据与数据中心的期望距离的方法。

像以前一样，标准差的计算方法是取方差的平方根，如下所示：

$$\sigma = \sqrt{Var(X)}$$

我们可以用和以前一样的符号表示标准差。

这就是说，老虎机收益的标准差是$\sqrt{2.6971}$,即1.642，这表示从平均情况看来，我们的每一局收益与期望收益-0.77之间的距离是1.642。

动动脑

你愿意老虎机的方差高一些还是低一些？为什么？

世上没有傻问题

问： 这么说期望与均值极为相似，那么对于概率分布来说，有没有类似中位数或是众数之类的东西呢？

答： 你可以算出最可能出现的概率，这就有点儿像众数，但一般不需要这么做。在研究概率分布的时候，统计师最感兴趣的测量值就是期望。

问： 期望是不是应该等于X能够取用的某个数值？

答： 不一定。就像一个数据集的均值不一定等于这个数据集中的某个数据，一个概率分布的期望也不一定等于X能够取用的一个数值。

问： 这里的方差和标准差和我们以前研究过的数值的方差和标准差是一样的吗？

答： 是一样的，不过这一次研究的是概率分布。数据集的方差和标准差是量度数据与均值的距离的方法，而概率分布的方差和标准差是量度一些特定数值的概率的分散情况的方法。

问： 我觉得$E(X-\mu)^2$很容易让人混淆，这个算式是不是等于求出$E(X-\mu)$再求平方？

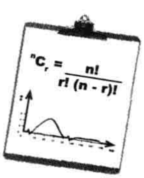
答： 不对，这是两个不同的算式。$E(X-\mu)^2$表示先求所有结果的平方，再求期望；如果先求出$E(X-\mu)$，再将结果平方，就会得出截然不同的答案。

从技术上说，你算的是$E((X-\mu)^2)$，但通常不这么写。

问： 那么方差低的老虎机和方差高的老虎机有何区别？

答： 方差高的老虎机表示你的整体收益变化大得多，整体上的赢钱数额更不可预期。

一般说来，方差越小，每一局的平均收益就越接近期望值。老虎机的方差越大，整体收益的可靠性越低。

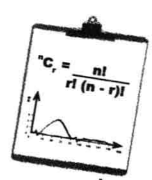


重要统计量

期望

“变量X的期望”计算公式如下：

$$E(X) = \sum xP(X=x)$$



重要统计量

方差

方差计算公式如下：

$$Var(X) = E(X-\mu)^2$$



下面是随机变量X的概率分布：

x	1	2	3	4	5
P(X = x)	0.1	0.25	0.35	0.2	0.1

1. E(X)是多少?

2. Var(X)是多少?

练习解答

下面是随机变量X的概率分布：

x	1	2	3	4	5
P(X = x)	0.1	0.25	0.35	0.2	0.1

1. E(X)是多少？

$$E(X) = \sum xP(X=x)$$

每个数值与其发生概率相乘，然后将所有乘积求和。

$$= 1\times0.1 + 2\times0.25 + 3\times0.35 + 4\times0.2 + 5\times0.1$$

$$= 0.1 + 0.5 + 1.05 + 0.8 + 0.5$$

$$= 2.95$$

2. Var(X)是多少？

逐个取用x，算出$(x-\mu)^2$。将结果与x的发生概率相乘，最后将全部乘积相加。

$$Var(X) = E(X-\mu)^2$$

$$= \sum (x-\mu)^2 P(X=x)$$

$$= (1-2.95)^2\times0.1 + (2-2.95)^2\times0.25 + (3-2.95)^2\times0.35 + (4-2.95)^2\times0.2 + (5-2.95)^2\times0.1$$

$$= (-1.95)^2\times0.1 + (-0.95)^2\times0.25 + (0.05)^2\times0.35 + (1.05)^2\times0.2 + (2.05)^2\times0.1$$

$$= 3.8025\times0.1 + 0.9025\times0.25 + 0.0025\times0.35 + 1.1025\times0.2 + 4.2025\times0.1$$

$$= 0.38025 + 0.225625 + 0.000875 + 0.2205 + 0.42025$$

$$= 1.2475$$





案件：不断变化的期望

统计邦播放过许多大家喜闻乐见的智力竞赛节目，其中有一个节目叫做“明与暗”，规则是这样的：向参赛者出示几个盒子，每个盒子里装有不同数额的钱，参赛者必须选择一个盒子，但不能看盒子里面有什么。剩下的盒子会一个接一个打开，每打开一个盒子，参赛者都有机会进行选择：留下原先选择的盒子中的钱（不能看），或根据装在其余未打开的盒子里的钱的总额另得一份奖金。根据参赛者得到的奖金，统计邦海豹保护区亦会得到一笔捐款。

最近的一位参赛者是一名业余统计师，他看出只要知道所有盒子的期望，就能增加胜算。他刚刚算完期望，制片人就来了。

“再过三分钟你就该上场了”，制片人说，“我们改过所有盒子里的数额了，和原来相比，现在的金额差10美元就翻倍。”

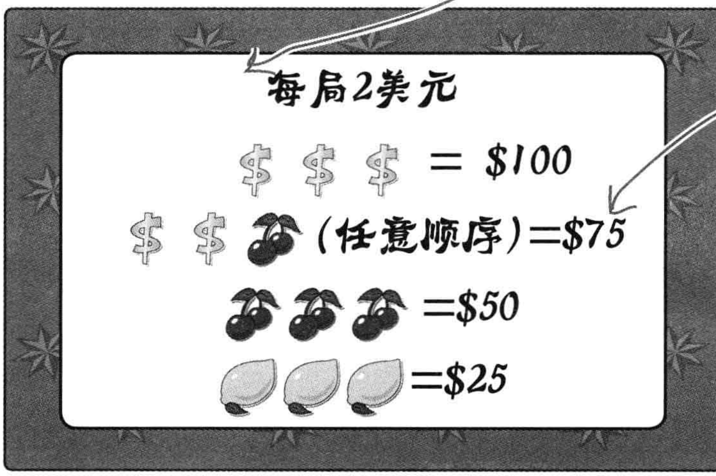
参赛者惊慌失措地瞪着制片人，难道他的全部计算都泡汤了吗？他不可能在三分钟以内从头算出期望。他该怎么办？

这位参赛者如何才能以前所未有的速度算出新的期望？

肥蛋改了价码

几分钟前，肥蛋改了老虎机的赌本和奖金，下面是新价码：

从每局1美元涨到每局2美元。



赢金是原赢金的5倍。



老虎机每一局（拉一次杆）的赌本现在从1美元变成了2美元，而赢金翻了5倍。要是赢了，就能捞更多钱了。

下面是新概率分布。

y	-2	23	48	73	98
P(Y = y)	0.977	0.008	0.008	0.006	0.001

这一次我们不用X，而用Y。



只要算出期望和方差，就能知道长期的收益情况。

动动笔

新概率的方差和期望是多少？这些数值与之前的收益分布期望-0.77和方差2.6971相比如何？

y	**-2**	**23**	**48**	**73**	**98**
P(Y = y)	0.977	0.008	0.008	0.006	0.001

动动笔解答

新概率的方差和期望是多少？这些数值与之前的收益分布期望-0.77和方差2.6971相比如何？

y	-2	23	48	73	98
P(Y = y)	0.977	0.008	0.008	0.006	0.001

$$E(Y) = (-2) \times 0.977 + 23 \times 0.008 + 48 \times 0.008 + 73 \times 0.006 + 98 \times 0.001$$

$$= -1.954 + 0.184 + 0.384 + 0.438 + 0.098$$

$$= -0.85$$

$$Var(Y) = E(Y - \mu)^2$$

$$= \sum (y - \mu)^2 P(Y=y)$$

$$= (-2+0.85)^2 \times 0.977 + (23+0.85)^2 \times 0.008 + (48+0.85)^2 \times 0.008 + (73+0.85)^2 \times 0.006 +$$

$$(98+0.85)^2 \times 0.001$$

$$= (-1.15)^2 \times 0.977 + (23.85)^2 \times 0.008 + (48.85)^2 \times 0.008 + (73.85)^2 \times 0.006 + (98.85)^2 \times 0.001$$

$$= 1.3225 \times 0.977 + 568.8225 \times 0.008 + 2386.3225 \times 0.008 + 5453.8225 \times 0.006 +$$

$$9771.3225 \times 0.001$$

$$= 1.2920825 + 4.55058 + 19.09058 + 32.722935 + 9.7713225$$

$$= 67.4275$$

期望稍微下降了一点儿，因此从长期看来，我们每局可望赔0.85美元；方差增大，这表示从长期看来，我们有可能在这台老虎机上赔更多的钱，但确定性更小。



新旧收益互有关联。

每一局的赌本上涨到2美元，赢金则是原来的5倍。由于新旧收益之间存在关系，所以，也许它们的期望和方差也存在关系。

让我们找出这种关系。

奇妙池



现在是代数时间。你的**任务**是将一些数字从奇妙池里捞出来，将它们放入计算式中的空白位置。每个数字**只能**用一次，但不需要把所有数字都用上。**目标**：根据老虎机的旧收益表达式得出新收益表达式。X代表旧收益，Y代表新收益。

X = (原收益) - (新赌本)

= (原收益) - ....

(原收益) = ... + ...

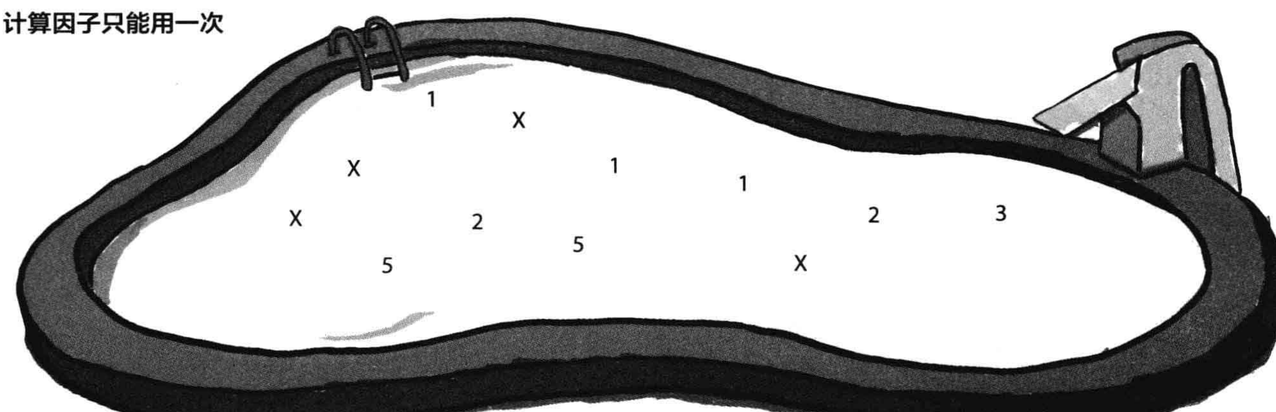
Y = 5 (原收益) - (新赌本)

= 5(............... + ...) - ....

= 5 .... + ... - ....

= .... + ...

注意：从池里捞出的每个计算因子只能用一次



奇妙池解答



现在是代数时间。你的**任务**是将一些数字从奇妙池里捞出来，将它们放入计算式中的空白位置。每个数字**只能**用一次，但不需要把所有数字都用上。**目标**：根据老虎机的旧收益表达式得出新收益表达式。X代表旧收益，Y代表新收益。

X = (原收益) - (新赌本)

= (原收益) - 1 ← 原赌本是1美元。

(原收益) = X + 1 ← 这个式子表示基于X的原收益。

我们可以将原收益表达式代入。

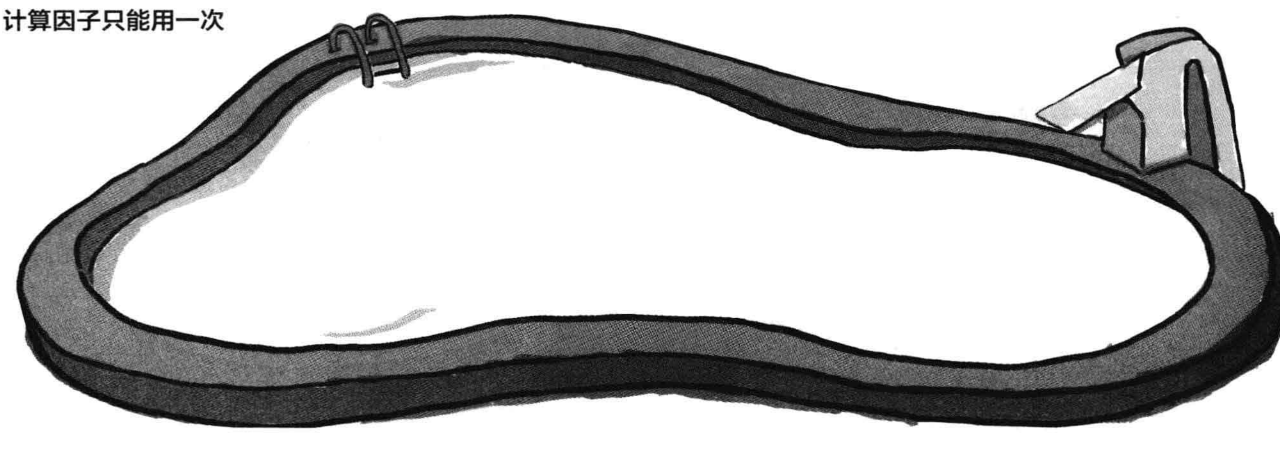
Y = 5 (原收益) - (新赌本)

= 5(X + 1) - 2

= 5 X + 5 - 2

= 5 X + 3 ← 所以Y=5X+3，这就是X与Y之间的确定关系。

注意：从池里捞出的每个计算因子只能用一次



E(X) 与E(Y)之间存在线性关系

我们发现，新收益与原收益可以通过Y = 5X + 3联系起来，其中，Y为新收益，X为原收益。现在我们要看看E(X)与E(Y)之间以及Var(X)与Var(Y)之间是否存在某种关系。

如果存在某种关系，我们就能在肥蛋改价码时大大节省计算新期望和新方案的时间。只要知道新结果和原结果之间的关系，我们就能迅速算出新期望和新方差。

动动笔

让我们看看E(X)与E(Y)的关系以及 Var(X)与 Var(Y)的关系是否有某种固定模式。

1. E(X)等于-0.77，E(Y) = -0.85，5 x E(X)是多少？ 5 x E(X) + 3是多少？结果与E(Y)有何关系？

2. Var(X) = 2.6971，Var(Y) = 67.4275，5 x Var(X)是多少？ 52 x Var(X)是多少？ 结果与Var(Y)有和关系？

3. 如何将这种关系推广至所有Y = aX + b的概率分布？

动动笔解答

让我们看看E(X)与E(Y)的关系以及Var(X)与Var(Y)的关系是否有某种固定模式。

1. E(X)等于-0.77，E(Y) = -0.85，5 x E(X)是多少？ 5 x E(X) + 3是多少？结果与E(Y)有何关系？

   $5 \times E(X) = -3.85$

   $5 \times E(X) + 3 = -0.85$

   $E(Y) = 5 \times E(X) + 3.$

2. Var(X) = 2.6971，Var(Y) = 67.4275，5 x Var(X)是多少？ 52 x Var(X)是多少？ 结果与Var(Y)有和关系？

   $5 \times Var(X) = 13.4855$

   $5^2 \times Var(X) = 67.4275$

   $Var(Y) = 5^2 \times Var(X)$

3. 如何将这种关系推广至所有Y = aX + b的概率分布？

   $E(aX + b) = a E(X) + b$

   $Var(aX + b) = a^2 Var(X)$

老虎机变换



你在前几页完成了哪些工作？

首先，你求出X的期望与方差，这里的X代表你在每一局中有望获得的收益。

然后，你想知道肥蛋的价格变化会造成什么结果，但不想完全从头开始计算期望与方差，于是你算出新收益与原收益之间的关系，再利用这种关系计算新期望与新方差。得出：

$$E(5X + 3) = 5E(X) + 3$$

$$Var(5X + 3) = 5^2Var(X)$$

线性变换的通用公式

我们可以将以下公式推广至任意随机变量，若随机变量为X：

期望乘以a，然后加b

$$E(aX + b) = aE(X) + b$$

取a的平方，乘以X的方差（忽略b）

$$Var(aX + b) = a^2Var(X)$$

这就是所谓的线性变换，因为X发生的是线性变化——即基础概率保持不变，但数值变为新值，其形式为：aX+b。

世上没有傻问题

问： a和b必须是常数吗？

答： 是的，如果a和b是变量，那么以上结果不成立。

问： 方差中的b哪里去了？

答： 在概率分布中增加一个常数仅对期望有影响，对整个方差没有影响。

在变量中增加一个常数不过是将概率分布移动一下，分布的形状依然不变。也就是说，期望以b为幅度进行偏移，但由于形状保持不变，所以方差也保持不变。

问： 我很惊奇，方差会乘以一个a^2，这是为什么？

答： 变量乘以一个常数意味着所有基础数据都乘以该常数。

在计算方差的过程中要计算各基础数据的平方。由于基础数据都乘以a，因此最终结果是方差乘以a^2。

问： 我必须记住如何做线性变换吗？这重要吗？

答： 是的，很重要。从长远看这能为你节省时间，不必数据一发生变化，你就得从头计算概率分布的期望和方差。相反，你可以将已经算得的期望和方差代入上式，从而得出新概率分布的期望和方差。

懂得做线性变换还可以帮助你考场得意，首先，知道简便算法可以帮助你节约时间；另外，考卷上不一定会给出基础概率分布，你的已知条件可能是变量的期望，你可能必须根据最基本的信息对其进行变换。

问： 我从头到尾算出了期望和方差，结果却是错的，这是为什么？

答： 你现在知道了吧，计算期望和方差是很容易出错的。如果按照常规算法，很容易不是这里错，就是那里错。尽量使用统计简化算法，这样效果会好一些。

破案：不断变化的期望

前面那位参赛者如何才能以前所未有的速度算出新的期望?

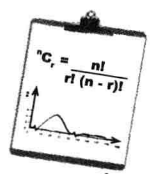


参赛者惊慌失措地左顾右盼了一会儿，接着释然了—— 数值的变化毕竟不是什么大问题。

参赛者已经花了一些时间算出所有盒子中的原有数值的期望，并由此获知有多少钱在向他招手。

制片人已经告诉过他，新奖金比原奖金的2倍少10美元，也就是说，这是一个线性变换。如果用X代表原奖金，用Y代表新奖金，则数值变换形式为：$Y = 2X-10$。

参赛者用$E(2X-10) = 2E(X)-10$求出$E(Y)$，也就是说，只要将原期望翻倍，再减去10，就能求出新期望。



重要统计量

线性变换

如果你有一个变量X，同时还有数字a和b，则：

$$E(aX + b) = aE(X) + b$$

$$Var(aX + b) = a^2Var(X)$$

要点

- 概率分布描述了一个给定变量的所有可能结果的概率。
- 期望即所期望的长期平均结果，以$E(X)$或μ表示，计算式为$E(X) = \Sigma xP(X=x)$.
- X的函数的期望为：$E(f(X)) = \Sigma f(x)P(X=x)$
- 概率分布的方差算式为：$Var(X) = E(X - \mu)^2$
- 概率分布的标准差算式为：$\sigma = \sqrt{Var(X)}$
- 当变量X按照aX+b的形式发生变换（其中a和b都是常数），则为线性变换，其方差和期望计算式为：

  $E(aX + b) = aE(X) + b$

  $Var(aX + b) = a^2Var(X)$



这么说如果我想多玩几种赌博游戏，通过线性变换能迅速算出期望和方差？

使用线性变化和多玩几种赌博游戏有区别

进行线性变换后，所有的概率都保持不变，但可能出现的数值发生变化——发生变换的是数值而非概率。这些可能数值的数目仍然不变。

如果多玩几种其他游戏，则数值和概率都发生变化，就连可能数值的数目也会发生变化。这时不可能只对数值进行转化，而概率的计算会迅速变得错综复杂。

让我们看一个简单的实例。假设你在玩一台非常简单的老虎机，概率分布为X。

x	**-1**	**5**
P(X = x)	0.9	0.1

为了求出2X的概率分布，只需将X乘以2，由于潜在收益翻倍，因此基础数据发生了变化。

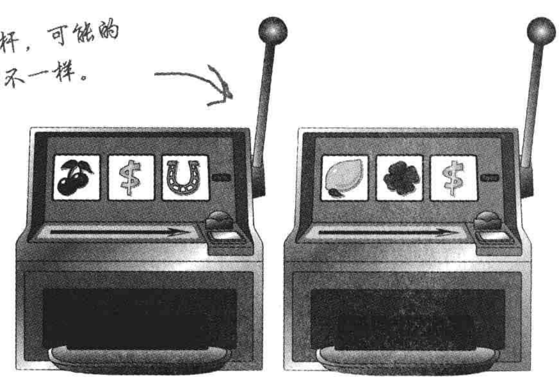
这里的数值乘以2，概率保持不变。



2x	**-2**	**10**
P(2X = 2x)	0.9	0.1

如果想在这台老虎机上玩两局，结果会如何呢？你需要从头开始计算概率分布，这时要考虑两局赌局可能出现的所有结果。

相当于拉两次杆，可能的收益和概率都不一样。



如果两局输则y=-2。

w	**-2**	**4**	**10**
P(W = w)	0.81	0.18	0.01

W代表两局赌局的结果。

两局都赢的话y=10。

一局-1，一局5，则y=4。

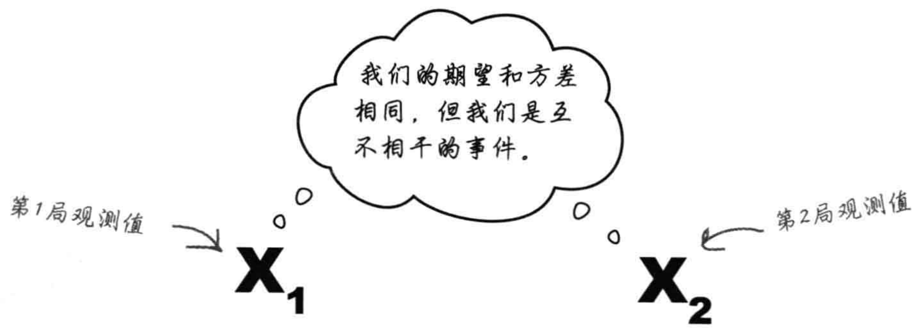
这一次概率和数值都变了，那么我们该如何求出这种情况的期望与方差？

每一次拉杆为一个独立观测值

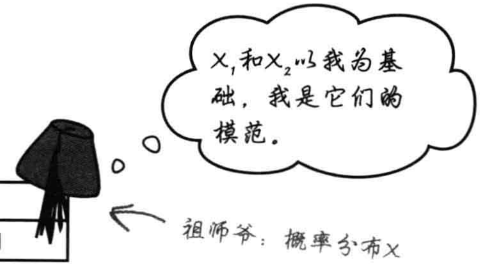
在赌博机上连玩多局赌局时，每一局称为一个事件，每一局的结果称为一个**观测值**。每一个观测值具有相同的期望和方差，但观测值互有差别，不可能每一局的收益都一样。

每一局赌局称为一个事件，每一局赌局的结果称为一个观测值

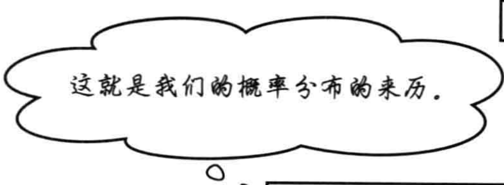
我们需要用某种办法对不同赌局或观测值进行区分，如果用X代表老虎机收益的概率分布，则把第一个观测值称为X_1，把第二个观测值称为X_2。



X_1和X_2都具有和X一样的概率、可能值、期望和方差，也就是说，虽然它们是互不相干的观测值，并且结果也不同，但它们的概率分布相同。



x	-1	5
P(X = x)	0.9	0.1



x_1	-1	5
$P(X_1 = x_1)$	0.9	0.1

x_2	-1	5
$P(X_2 = x_2)$	0.9	0.1

我们希望求出两局老虎机赌局的期望和方差，实际上就是要求X_1+X_2的期望和方差，让我们看一些快速算法：

观测值速算法

让我们求出$X_1 + X_2$的期望和方差。



$X_1 + X_2$ 并不等于 2X。

$X_1 + X_2$ 表示你在考虑X的两个观测值，2X表示你有一个观测值，但其可能数值翻倍。

期望

首先算 $E(X_1 + X_2)$.

$$
\begin{aligned}
E(X_1 + X_2) &= E(X_1) + E(X_2) \\
&= E(X) + E(X) \\
&= 2E(X)
\end{aligned}
$$

由于X_1和X_2的概率分布都沿袭X的概率分布，因此$E(X_1)$和$E(X_2)$都等于E(X)。

换句话说，如果我们已知两个观测值的期望，则将E(X)乘以2即可。即，如果要在E(X)=-0.77的老虎机上玩两局，则相应期望为-0.77 × 2=-1.54。

我们可以将整个结论推广至多个观测值，若我们想求出n个观测值的期望，则可按下式计算：

如果有n个观测值，则用E(X)乘以n即可。

$$E(X_1 + X_2 + \cdots X_n) = nE(X)$$

方差

那么$Var(X_1 + X_2)$又如何计算呢？下面是计算方法：

$$
\begin{aligned}
Var(X_1 + X_2) &= Var(X_1) + Var(X_2) \\
&= Var(X) + Var(X) \\
&= 2Var(X)
\end{aligned}
$$

由于X_1和X_2沿袭X的概率分布，因此$Var(X_1)$和$Var(X_2)$与Var(X)相同。

也就是说，如果我们在Var(X) = 2.6971的老虎机上玩两局，则方差为2.6971 × 2= 5.3942。

我们可以将整个结论推广至任何数目的独立观测值。如果有X的n个独立观测值，则：

Var(X)乘以观测值的数目n。

$$Var(X_1 + X_2 + \cdots X_n) = nVar(X)$$

也就是说，为了求出多个观测值的期望和方差，只要用观测值的数目乘以E(X)和Var(X)就行了。

世上没有傻问题

问： 难道$E(X_1+X_2)$与$E(2X)$不一样？

答： 看似相似，其实不然，它们是两个概念。

如果是$E(2X)$，则表示你想将一个变量的基础数据翻倍，然后求其期望和方差。也就是说，变量只有一个，但数值变为两倍。

如果是$E(X_1+X_2)$，则表示你观测到了X的两个独立结果，需要求其综合期望。例如，如果X代表一局赌局的概率分布，则X_1+X_2代表两局游戏的概率分布。

问： 这么说X_1和X_2是一样的？

答： 它们的概率分布相同，但它们本身是不同的结果(或者说观测值)。例如，X_1可以指第1局，X_2指第2局，它们具有相同的概率分布，但实际结果可以不一样。

问： 我发现新方差是nVar(X)，而不是像线性变换的结果一样是$n^2Var(X)$，这是为什么？

答： 这一次我们有一系列的独立观测值，这些观测值都有相同的概率分布，于是我们可以将所有观测值的方差相加，求出整个方差，如果有n个独立观测值，则结果为nVar(X)。

在计算方差Var(nX)时，我们将基础数据乘以n，由于方差是通过取基础数据的平方得到的，因此所求方差为$n^2Var(X)$。

重要统计量

独立观测值

使用下列公式计算其方差：

$$E(X_1 + X_2 + ... + X_n) = nE(X)$$

$$Var(X_1 + X_2 + ... + X_n) = nVar(X)$$

要点

- 概率分布描述了一个给定随机变量的所有可能结果的概率。
- 一个随机变量X的期望等于我们所期望的长期平均值，以E(X)或μ表示。计算式为：

  $E(X) = \Sigma xP(X=x)$

- 一个随机变量X的方差计算式为：

  $Var(X) = E(X - \mu)^2$

- 标准差σ是方差的平方根。
- 当一个随机变量从X变换为aX+b时，则为线性变换，其中a和b均为常数。其期望和方差计算式为：

  $E(aX + b) = aE(X) + b$

  $Var(aX + b) = a^2Var(X)$

是线性变换，还是独立观测值？

下面是一系列实例，假定已知每个X的概率分布，你的任务是说出可以通过哪种方法解决各个问题：是线性变换，还是独立观测值？

	线性变换	独立观测值
一杯超大杯咖啡的咖啡量；X是普通杯咖啡的咖啡量。	☐	☐
每天多喝一杯咖啡；X是一杯咖啡的量。	☐	☐
求买10张彩票的净收益；X是买一张彩票的净收益。	☐	☐
求彩票价格上涨后每买一张彩票的净收益；X是买一张彩票的净收益。	☐	☐
多买一只母鸡，靠它下蛋做早餐；X是某个品种的鸡每周的产蛋量。	☐	☐

是线性变换，还是独立观测值？解答

下面是一系列实例，假定已知每个X的概率分布，你的任务是说出可以通过哪种方法解决各个问题：是线性变换，还是独立观测值？

	线性变换	独立观测值
一杯超大杯咖啡的咖啡量；X是普通杯咖啡的咖啡量。	☑	☐
每天多喝一杯咖啡；X是一杯咖啡的量。	☐	☑
求买10张彩票的净收益；X是买一张彩票的净收益。	☐	☑
求彩票价格上涨后每买一张彩票的净收益；X是买一张彩票的净收益。	☑	☐
多买一只母鸡，靠它下蛋做早餐；X是某个品种的鸡每周的产蛋量。	☐	☑

每买一张彩票的收益与是否购买其他彩票无关。

彩票价格改变则期望收益改变，但收益概率不变，因此可以通过线性变换解答。

练习

本地餐厅正在搞促销活动，每块糕饼售价0.50美元，并藏有一条神秘信息。大部分信息都不过是预祝购买者前程似锦，但还有一部分却表示可为晚餐打折。折扣2美元的概率是0.1，折扣5美元的概率是0.07，折扣10美元的概率是0.03。

如果X为顾客的净收益，那么X的概率分布如何？E(X)和 Var(X)等于多少？

餐厅决定将糕饼价格调高1美元，新的期望和方差是多少？



本地餐厅正在搞促销活动，每块糕饼售价0.50美元，并藏有一条神秘信息。大部分信息都不过是预祝购买者前程似锦，但还有一部分却表示可为晚餐打折。折扣2美元的概率是0.1，折扣5美元的概率是0.07，折扣10美元的概率是0.03。

如果X为顾客的净收益，那么X的概率分布如何？E(X)和 Var(X)等于多少？

下面是X的概率分布：

x	-0.5	1.5	4.5	9.5
P(X = x)	0.8	0.1	0.07	0.03

$$E(X) = (-0.5)\times0.8 + 1.5\times0.1 + 4.5\times0.07 + 9.5\times0.03$$

$$= -0.4 + 0.15 + 0.315 + 0.285$$

$$= 0.35$$

$$Var(X) = E(X - \mu)^2$$

$$= \sum(x - \mu)^2P(X=x)$$

$$= (-0.5-0.35)^2\times0.8 + (1.5-0.35)^2\times0.1 + (4.5-0.35)^2\times0.07 + (9.5-0.35)^2\times0.03$$

$$= (-0.85)^2\times0.8 + (1.15)^2\times0.1 + (4.15)^2\times0.07 + (9.15)^2\times0.03$$

$$= 0.7225\times0.8 + 1.3225\times0.1 + 17.2225\times0.07 + 83.7225\times0.03$$

$$= 0.578 + 0.13225 + 1.205575 + 2.511675$$

$$= 4.4275$$

餐厅决定将糕饼价格调高1美元，新的期望和方差是多少？

餐厅将糕饼价格调高了0.5美元，即新的净收益模型为X - 0.5：

$$E(X - 0.5) = E(X) - 0.5$$

$$= 0.35 - 0.5$$

$$= -0.15$$

$$Var(X - 0.5) = Var(X)$$

$$= 4.4275$$

新老虎机在等你



肥蛋赌场买进一台新式老虎机，赌本更大，奖金更高。
下面是这台新老虎机的概率分布：

每一局赌本比其他老虎机更高，不过，看看奖金吧！

x	-5	395
P(X = x)	0.99	0.01

我们已经讲过单玩一台老虎机的期望和方差，也讲过在同一台老虎机上连玩几局的期望和方差，那么，要是在两台老虎机上玩两局呢？

在这种情况下，两台老虎机有两种各自独立、互不相同的概率分布：

x	-5	395
P(X = x)	0.99	0.01

这是肥蛋赌场新老虎机的当前收益。

y	-2	23	48	73	98
P(Y = y)	0.977	0.008	0.008	0.006	0.001

这是原来的老虎机的当前收益。

我们该怎么求在两台老虎机上各玩一局的期望和方差呢？

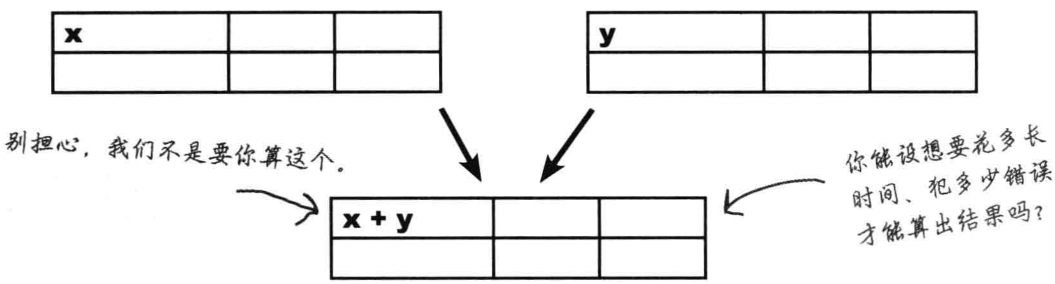


我们可以求出X+Y的概率分布，不过那样太费时间，而且有可能出差错。我在想有没有别的捷径可走？

E(X) + E(Y) = E(X + Y)

我们希望求出在每台老虎机上各玩一局的期望和方差，即希望求出E(X + Y)和Var(X + Y)，其中X和Y为代表两台老虎机的随机变量，X和Y相互独立。

实现此目的的一个方法是算出X+Y的概率分布，然后计算期望和方差。



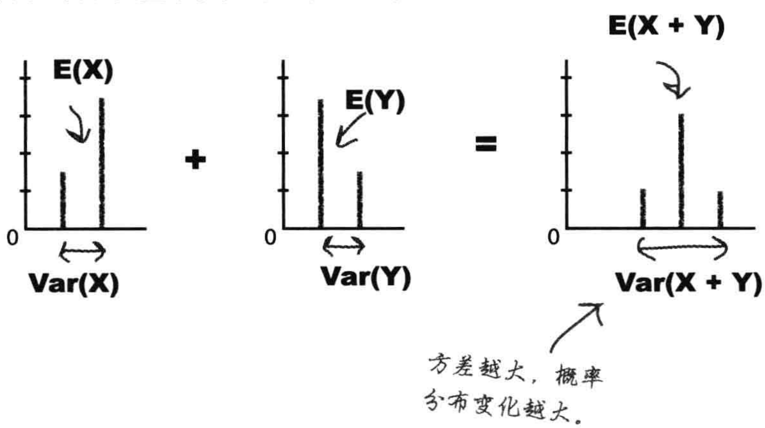
幸亏我们不必这么做。只要将E(X)和E(Y)相加，就能求出E(X + Y)。

E(X + Y) = E(X) + E(Y)

意义显而易见，例如，如果你玩两局，一局有望赢5美元，另一局有望赢10美元，则总体上有望赢5美元+10美元=15美元。

Var(X + Y) = Var(X) + Var(Y)

类似地可以求出方差，只要将两个方差相加即可。对于所有独立随机变量来说，这些结论全都成立。





方差加法仅适用于独立随机变量

如果X和Y相互不独立，则Var(X+Y)不再等于Var(X) + Var(Y)。

E(X)-E(Y)= E(X - Y)

随机变量不仅能相加，还能相减，这时不是X+Y，而是X-Y。

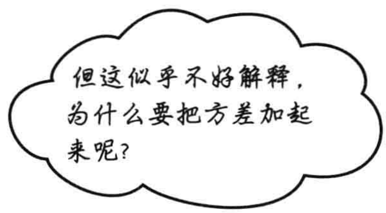
如果面对的是两个随机变量的差，就很容易求出期望E(X-Y)，只要用E(X)减去E(Y)即可。

X-Y的方差Var(X-Y)则不那么直观——为了求Var(X-Y)，需要将两个方差加起来。

$$E(X - Y) = E(X) - E(Y)$$

$$Var(X - Y) = Var(X) + Var(Y)$$

要将方差相加，小心哦！





这是因为变异性增大了。

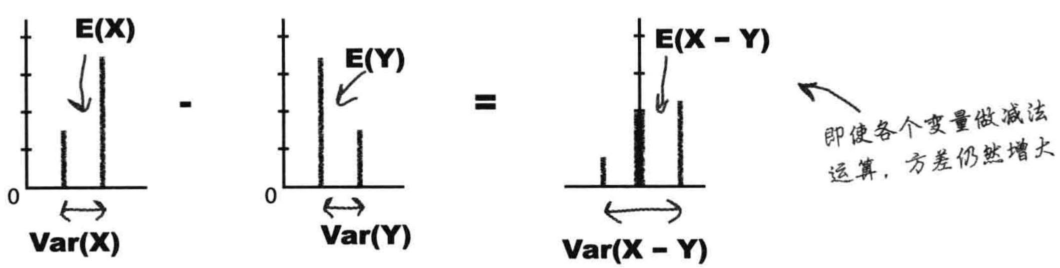
若我们用一个随机变量减另一个随机变量，概率分布的方差依然增大。



若将两个随机变量相减，则方差要相加。

猛一看，这个算法有违直观，因此很容易搞错。切记：如果两个变量是独立变量，则

Var(X - Y) = Var(X) + Var(Y)



将两个相互独立的随机变量相减后的方差与将两个变量相加后的方差是一模一样的，变异性只会增加，不会减少。

独立随机变量做减法运算，方差依然增大。

线性变换也可以做加减运算

事情还没有结束，像随机变量加减运算一样，线性变换也可以做加减运算。

假设出现这种情况：肥蛋赌场更改了两台老虎机（甚至只是其中一台老虎机）的赌本和奖金，我们最后需要做的是，算出整个概率分布，以便求出新的方差和期望。

真走运，我们可以用另一种简便算法。

假设X和Y老虎机的收益变了，使得X的收益为aX，Y的收益为bY，其中a和b为任意数字。

为了求出aX和bY这两个组合的期望和方差，可以使用以下简便算法：

X → aX

Y → bY

a和b可为任意数字。

aX与bY相加

为了求出aX + bY的期望和方差，可使用下列算式：

$$E(aX + bY) = aE(X) + bE(Y)$$

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y)$$

如前所述，由于是线性变换，所以取数字的平方。

这是线性变换，所以这里用平方。

aX与bY相减

若将随机变量相减并计算E(aX－bY)和Var(aX－bY)，
可使用下列算式：

$$E(aX - bY) = aE(X) - bE(Y)$$

$$Var(aX - bY) = a^2Var(X) + b^2Var(Y)$$

如前所述，即使随机变量做减法，方差仍然做加法。

切记：将方差相加。

世上没有傻问题

问： 如果X和Y代表赌局，那么aX+bY是表示“a局X赌局+b局Y赌局”吗？

答： aX + bY其实是表示两个线性变换相加，换句话说，X和Y的基础数据变了，这与独立观测值不一样，对于独立观测值来说，每一局都是一个独立观测值。

问： 我看不出什么时候会用到X − Y。这能达到什么目的呢？

答： 在你希望求出两个变量的差时，X−Y的确十分有用。E(X−Y)有点儿像在说“你所期望的X与Y的差别”，而Var(X−Y)则指出方差。

问： 为什么把X − Y的方差加起来？你肯定应该做减法吧？

答： 猛一看这有违直觉，不过，当你用一个变量减另一个变量时，其实变异性是增大的，因此方差也增大。变量相减的变异性与变量相加的变异性其实是一样的。

还有一种理解方法：计算方差时会取基本数值的平方，Var(X+bY)等于Var(X)+b^2Var(Y)，如果b＝-1，则得出Var(X−Y)，由于$(-1)^2=1$，因此Var(X−Y) = Var(X) + Var(Y)。

问： 如果X和Y相互不独立，还能这么计算吗？

答： 不行，只有在X和Y相互独立时才能这么做，如果要求相关的X +Y的方差，则必须从头计算概率分布。

问： 似乎$X_1 + X_2$的规律也同样适用于X + Y，对吗？

答： 对的，只要X、Y、X_1及X_2相互独立就行。

要点

- **X的独立观测值**与X不同，每个观测值都具有相同的概率分布，但结果各不一样。
- 如果X1，X2，...，Xn是X的独立观测值，则：

  $E(X_1 + X_2 + ... + X_n) = nE(X)$

  $Var(X_1 + X_2 + ... X_n) = nVar(X)$

- 如果X和Y是独立随机变量，则：

  E(X + Y) = E(X) + E(Y)

  E(X − Y) = E(X) − E(Y)

  Var(X + Y) = Var(X) + Var(Y)

  Var(X − Y) = Var(X) + Var(Y)

- X和Y的线性变换的期望和方差用下列各式进行计算：

  E(aX + bY) = aE(X) + bE(Y)

  E(aX − bY) = aE(X) − bE(Y)

  Var(aX + bY) = a2Var(X) + b2Var(Y)

  Var(aX − bY) = a2Var(X) + b2Var(Y)

练习

下表中有一些期望和方差，请写出其计算公式或简便算法，必要时假定变量为独立变量。

统计量	简便算法或公式
$E(aX + b)$	
$Var(aX + b)$	
$E(X)$	
$E(f(X))$	
$Var(aX - bY)$	
$Var(X)$	
$E(aX - bY)$	
$E(X_1 + X_2 + X_3)$	
$Var(X_1 + X_2 + X_3)$	
$E(X^2)$	
$Var(aX - b)$	



练习

某家餐厅备有两份菜单，一份是周末菜单，一份是平日菜单。每份菜单有四种定价，就餐者的消费概率分布如下：

平日

x	10	15	20	25
P(X = x)	0.2	0.5	0.2	0.1

周末

y	15	20	25	30
P(Y = y)	0.15	0.6	0.2	0.05

你会期望谁给餐厅带来最大营业额：周末20位用餐者，还是平日25位用餐者？

练习解答

下表中有一些期望和方差，请写出其计算公式或简便算法，必要时假定变量为独立变量。

统计量	简便算法或公式
$E(aX + b)$	$aE(X) + b$
$Var(aX + b)$	$a^2Var(X)$
$E(X)$	$\sum xP(X = x)$
$E(f(X))$	$\sum f(x)P(X = x)$
$Var(aX - bY)$	$a^2Var(X) + b^2Var(Y)$
$Var(X)$	$E(X - \mu)^2 = E(X^2) - \mu^2$
$E(aX - bY)$	$aE(X) - bE(Y)$
$E(X_1 + X_2 + X_3)$	$3E(X)$
$Var(X_1 + X_2 + X_3)$	$3Var(X)$
$E(X^2)$	$\sum x^2P(X = x)$
$Var(aX - b)$	$a^2Var(X)$



练习解答

某家餐厅备有两份菜单，一份是周末菜单，一份是平日菜单。每份菜单有四种定价，就餐者的消费概率分布如下：

平日

x	10	15	20	25
P(X = x)	0.2	0.5	0.2	0.1

周末

y	15	20	25	30
P(Y = y)	0.15	0.6	0.2	0.05

你会期望谁给餐厅带来最大营业额：周末20位用餐者，还是平日25位用餐者？

让我们先求出平日和周末的期望。X代表平日用餐者，Y代表周末用餐者。

$$E(X) = 10\times0.2 + 15\times0.5 + 20\times0.2 + 25\times0.1$$
$$= 2 + 7.5 + 4 + 2.5$$
$$= 16$$

$$E(Y) = 15\times0.15 + 20\times0.6 + 25\times0.2 + 30\times0.05$$
$$= 2.25 + 12 + 5 + 1.5$$
$$= 20.75$$

每一位用餐者是一个独立观测值，为了求出每一类用餐者的用餐金额，我们用期望乘以该类用餐者的数量。

25位用餐者在平日用餐，则：$25\times E(X) = 25\times16 = 400$

20位用餐者在周末用餐，则：$20\times E(Y) = 20\times20.75 = 415$

这说明，我们能够期望：20位周末用餐者支付的餐费高于25位平日用餐者支付的餐费。

发了！

通过学习本章你颇有斩获，你学会了用概率分布、期望、方差预测自己能在某台老虎机上赢多少钱。

你还发现了如何用线性变换和独立观测值预测在收益结构发生变化时或在同一台老虎机上多次赌博时有望赢得的奖金。



练习

山姆有两家喜欢去的餐厅，餐厅A一般比餐厅B贵，但食物品质一般好得多。

下面的两组概率分布描述了山姆在每家餐厅的消费意愿，一般说来，你觉得两家餐厅价格差别如何？差别的方差是多少？

餐厅A

x	**20**	**30**	**40**	**45**
P(X = x)	0.3	0.4	0.2	0.1

餐厅B

y	**10**	**15**	**18**
P(Y = y)	0.2	0.6	0.2



山姆有两家喜欢去的餐厅，餐厅A一般比餐厅B贵，但食物品质一般好得多。

下面的两组概率分布描述了山姆在每家餐厅的消费意愿，一般情况下，你觉得两家餐厅价格差别如何？差别的方差是多少？

餐厅A

x	20	30	40	45
P(X = x)	0.3	0.4	0.2	0.1

餐厅B

y	10	15	18
P(Y = y)	0.2	0.6	0.2

让我们先算X和Y的期望和方差。

$$E(X) = 20\times0.3 + 30\times0.4 + 40\times0.2 + 45\times0.1$$
$$= 6 + 12 + 8 + 4.5$$
$$= 30.5$$

$$Var(X) = (20-30.5)^2\times0.3 + (30-30.5)^2\times0.4 + (40-30.5)^2\times0.2 + (45-30.5)^2\times0.1$$
$$= (-10.5)^2\times0.3 + (-0.5)^2\times0.4 + 9.5^2\times0.2 + 14.5^2\times0.1$$
$$= 110.25\times0.3 + 0.25\times0.4 + 90.25\times0.2 + 210.25\times0.1$$
$$= 33.075 + 0.1 + 18.05 + 21.025$$
$$= 72.25$$

$$E(Y) = 10\times0.2 + 15\times0.6 + 18\times0.2$$
$$= 2 + 9 + 3.6$$
$$= 14.6$$

$$Var(Y) = (10-14.6)^2\times0.2 + (15-14.6)^2\times0.6 + (18-14.6)^2\times0.2$$
$$= (-4.6)^2\times0.2 + 0.4^2\times0.6 + 3.4^2\times0.2$$
$$= 21.16\times0.2 + 0.16\times0.6 + 11.56\times0.2$$
$$= 4.232 + 0.096 + 2.312$$
$$= 6.64$$

X和Y的差可以用模型X-Y表示。

$$E(X - Y) = E(X) - E(Y)$$
$$= 30.5 - 14.6$$
$$= 15.9$$

$$Var(X - Y) = Var(X) + Var(Y)$$
$$= 72.25 + 6.64$$
$$= 78.89$$

6 排列与组合

排序、排位、排



顺序有时很重要。

一一清点某些事物的**所有可能排序方法**耗时颇巨，可这却是计算某些概率**必不可少**的过程 —— 麻烦就在这里。在本章中，我们将介绍推导出这类信息的**简便方法**，为你免除清点一切可能结果的烦恼。来吧，让我们看看如何**计算概率**。

统计邦德比杯马赛



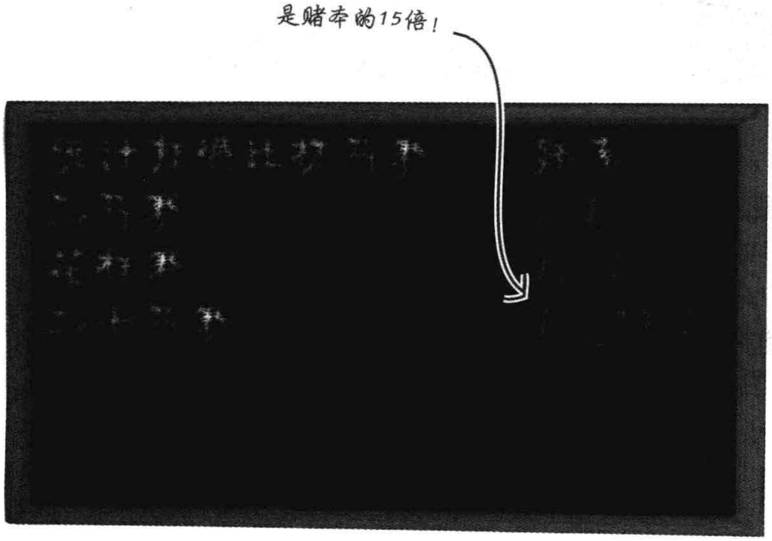
统计邦德比杯马赛是统计邦最重要的一项体育赛事，来自四面八方的骑师和他们品种各异的爱马将在这里一较高下，你可以对比赛结果下注。要是能押中每场比赛的前三名，大把钞票就到手了。

开幕赛在新马之间进行，参加比赛的都是一些初次进军赛场的嫩马，因此，没有前期比赛的统计量用以预测马匹的表现。也就是说，你必须假定每一匹马都有相同的得胜几率，这可以归结为简单概率问题。

当天的第一场比赛是三马赛。比赛即将开始，德比马场开始接受下注。你从肥蛋赌场赢了500美元，正好可以在德比马场花掉。只要能押中三匹马的最终排名，赔率可达7：1，即赌本翻7倍：可获3500美元。

我们该下注吗？让我们先求出几个概率再做决定不迟。





三马赛正在进行

第一场比赛在三匹马之间展开，十分简单直接。一心赢大钱的你需要预测马匹的最终排名，下面是参加比赛的三匹马。



翠香



拉托



福福

动动笔

比赛结果有几种可能（假定没有平局且每一匹马都跑完比赛）？

押中正确结果的概率是多大？

计算该赌局的期望收益。

提示：求出每种结果的概率分布，然后计算期望。

动动笔解答

比赛结果有几种可能（假定没有平局且每一匹马都跑完比赛）？
押中正确结果的概率是多大？

计算该赌局的期望收益。

比赛结果有6种可能：

翠香，拉托，福福
翠香，福福，拉托
拉托，翠香，福福
拉托，福福，翠香
福福，翠香，拉托
福福，拉托，翠香

因此，押中正确排名的概率为1/6。

押上500美元赌本（赔率7：1）后可以期望得到的收益的概率分布为：

三马赛

x	-500	3,500
P(X = x)	0.833	0.167

$$E(X) = -500 \times 0.833 + 3{,}500 \times 0.167$$
$$= 168$$

每比一局这样的比赛，我们可以期望收入168美元。





确实，大多数比赛的参赛马匹都不止三匹。

我们需要找出一个便捷的方法，通过这个方法，无论参加比赛的马匹数目是多少，都能求出马匹的最终排名有多少种可能。

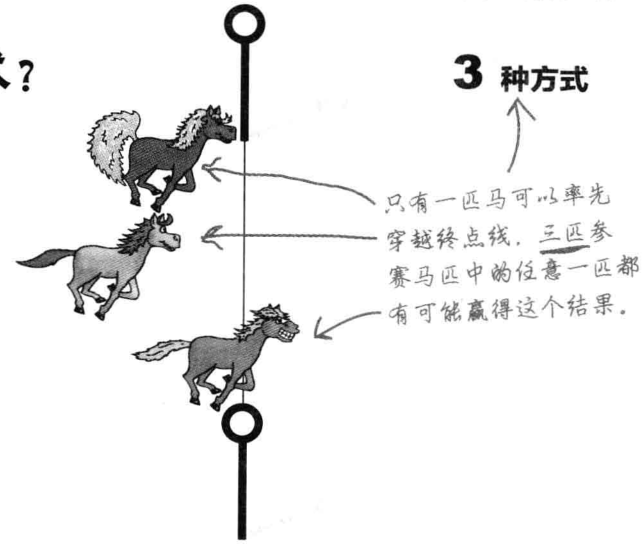
求三匹马最终排名状况的方法十分简单明了，因为只有6种可能局面。现在的麻烦在于，参加比赛的马匹越多，逐个写出最终排名的难度越大，所花费的时间越多。

让我们仔细观察参加比赛的三匹马的各种排名方式，看看是否有某种固定模式。为此我们可以一个一个地对名次进行考虑。

马儿们有几种穿越终点线的方式？

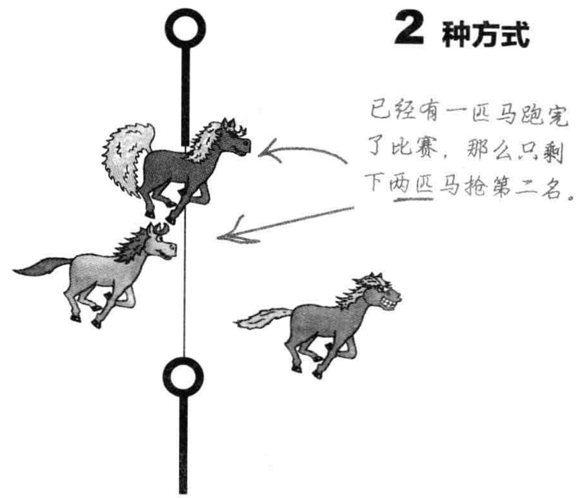
让我们先看第一名。

肯定有一匹马会成为冠军，三匹马中的任意一匹都有此可能。也就是说，占据第一名位置的方式有三种。

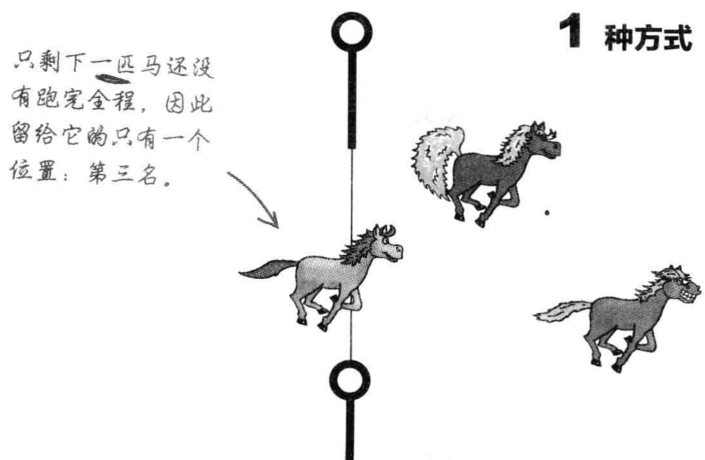


第二名是怎么个情况呢？

如果已经有一匹马跑完了比赛，那么还剩下两匹马，其中之一会成为第二名。即，占据第二名位置的方式有二种，这与跑第一名的马匹无关。



当有两匹马跑完比赛后，只剩下一个位置留给最后一匹马——第三名。



这对我们计算所有可能出现的最终排名有何帮助呢？

计算排位数目

前面讲到，第一名有三种占据方式，每一种方式对应着两种第二名的占据方式，无论前两名由谁占据，最后一名都仅有一种占据方式。即，三个位置的占据方式共计：

$$3 \times 2 \times 1 = 6$$

第一名的占据方式有3种。

第二名的占据方式有2种。

第三名的占据方式有1种。

3个位置共有6种占据方式。

这表示，我们不用把具体排名情况列举出来就可以做出结论：这3匹马有6种排名方式。

如果有n匹马呢？

我们已经讲过，3匹马共有3 × 2 × 1种排名方式，将这个算法推而广之，可以知道任意数目n的排名方式。即，如要算出n个独立对象的排名方式，可按下式进行计算：

$$n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

如此一来，不用一一列举每种可能的现象，也能算出n个独立对象的排名方式的确切数目。

这种计算方式称为一个数的**阶乘**，其数学表达式是感叹号，例如，3的阶乘写作3!，n的阶乘写作n!，读作“n的阶乘”。

因此，当我们写下n!，就表示“从大到小取n到1的所有数，并将这些数相乘”，即执行下列计算：

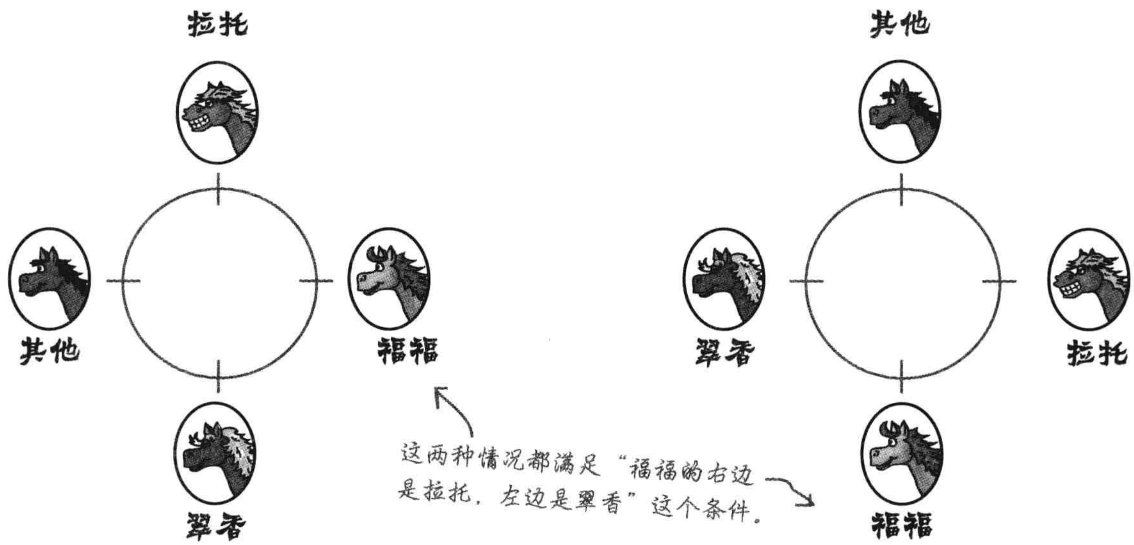
$$n! = n \times (n-1) \times (n-2) \times \cdots \times 3 \times 2 \times 1$$

许多计算器都将n!作为一个函数，这是使用n!的好处。例如，当要计算4个独立对象的排名方式的数目时，只需计算4!，即4 × 3 × 2 × 1 = 24。

圆形排位

前面讲到的计算规则有一个例外，那就是圆形排位。

下面举个例子。假定你想让4匹马围成一圈，并要求出可能的排位方式的数目。现在让我们看看这种情况：福福的右边是拉托，左边是翠香，符合这个要求的排位方式共有4种，下面是其中两种。



猛一看，这两种排位不一样，但其实呢，却是一样的。马与马的相对位置完全一样，唯一的区别是，第二幅图中的马儿们绕着圆圈动了一动。这就是说，马匹的某些排位方式实际上是完全一样的。

这一类问题该怎么解决呢?

关键是把其中一匹马的位置固定下来，比如福福。只要福福站在某个位置上不动，就能计算其余3匹马的排位方式，这样就能避免重复计算，得出正确的结果。

通常，如果有n个对象需要进行圆形排位，则可能的排位数目按下式进行计算：

$$(n-1)!$$

n个对象呈圆形排位的可能方式的数目。

世上没有傻问题

问： n!怎么读？

答： 读作“n的阶乘”。感叹号代表一种数学运算，和感情没有什么关系。

问： 阶乘只在排位物体的时候有用吗？

答： 绝对不是这样，阶乘在其他数学分支中也能派上用场，例如微积分，总的说来，这是十分有用的数学简便算法，只要进行这类乘法运算，就能看到阶乘符号。

阶乘符号的意思是“从大到小取n到1的所有数，并将这些数相乘”。

问： 如果n的数值是0呢？0的阶乘怎么求？

答： 0！为1，这个结果似乎有些奇怪，不过可以理解为“0个对象只有1种排列方法”。

问： 要是想求负数的阶乘该怎么办？或者非整数的阶乘该怎么求？

答： 阶乘仅针对正整数，因此无法求负数或非整数的阶乘。

可以这样理解，对零碎对象进行排位并无意义，你为之排位的每一个对象都被认定为一个完整的对象，同时，对象个数不可能是负数。

问： 阶乘的计算结果会是奇数吗？

答： 只有两种情况：在n为0或n为1时，n!＝1。

除此以外，所有其他数的阶乘均为偶数，这是因为，只要n大于等于2，计算式中就必定会包含2这个数字，2与任何整数相乘结果均为偶数。所以说，只要n大于等于2，n!均为偶数。

问： 计算大数的阶乘似乎是一种折磨，如果要求10!，就必须将10个数字相乘（10×9×8×7×6×5×4×3×2×1），结果会是一个很大的数。有没有简单点的办法？

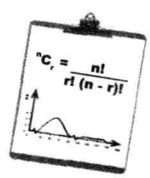
答： 有啊，许多科学计算器和绘图计算器都有阶乘按键（一般标有n!），你可以用这个按键进行计算。

问： 计算n个对象的圆形排位时，结果为（n－1）!。如果把顺时针和逆时针排位视为同一种情况进行计算，结果如何？

答： 如果这样的话，排位方式的数目则是(n－1)!/2。（n－1）！既考虑了顺时针的情况，也考虑了逆时针的情况，因此是实际要计算的结果的两倍，除以2就解决问题了。

问： 如果将对象呈圆形排位，且考虑对象的绝对位置，结果如何？

答： 这样的话，排位方式的数目为n!，这正好等于n个对象的排位方式的数目。



重要统计量

排位方式的计算公式

如果要求n个对象的可能排位方式的数目，则计算：

$n! = n \times (n-1) \times \cdots \times 3 \times 2 \times 1$

也就是说，将从n到1的数字全部相乘。

如果n个对象作圆形排位，则可能的排位方式的数目为$(n-1)!$。

练习

宝娜想给统计邦健身俱乐部打电话，但她的记性实在太差，她只知道电话号码由1、2、3、4、5、6、7组成，却忘记了顺序。她随机拨对号码的概率是多大？

有人提醒宝娜，电话号码的前3位是1、2、3的某种排位，后4位是4、5、6、7的某种排位。但她忘记了顺序，这时她拨对电话号码的概率有多大？

提示：这一次需要对两组数据作排位。

动动笔

统计邦德比马场要在本季末组织一次队列表演，马匹将沿着赛道排成圆形队列。马匹的确切顺序将随机抽取，你要是能猜中这个顺序，将会获得一笔奖金。

你猜中马匹列队顺序并获得奖金的概率是多大？

练习解答

宝娜想给统计邦健身俱乐部打电话，但她的记性实在太差，她只知道电话号码由1、2、3、4、5、6、7组成，却忘记了顺序。她随机拨对号码的概率是多大？

有7个数字，因此有7！种可能的排位方式。$7! = 7 \times 6 \times 5 \times 4 \times 3 \times 2 \times 1 = 5040$。

因此拨对号码的概率为$1/5040 = 0.0002$

有人提醒宝娜，电话号码的前3位是1、2、3的某个排位，后4位是4、5、6、7的某个排位。但她忘记了顺序，这时她拨对电话号码的概率有多大？

先将数字拆分为两组，第一组3个数字(1, 2, 3)，其余为第二组(4, 5, 6, 7)，得到：

1、2、3的排位方式的数目为$3! = 3 \times 2 \times 1 = 6$

4、5、6、7的排位方式的数目为$4! = 4 \times 3 \times 2 \times 1 = 24$

为了求出可能的排位方式的总数，可将两组排位结果的数目相乘，得到：

可能的排位方式的总数$3! \times 4! = 6 \times 24 = 144$

因此，拨对号码的概率为$1/144 = 0.0069$

动动笔解答

统计邦德比马场要在本季末组织一次队列表演，马匹将沿着赛道排成圆形队列。马匹的确切顺序将随机抽取，你要是能猜中这个顺序，将会获得一笔奖金。

你猜中马匹列队顺序并获得奖金的概率是多大？

10匹马作圆形队列，即马匹有9!种可能的顺序。

$9! = 362880$，即队列有362880种可能的顺序。

猜对结果的概率为$1/9!$，几乎等于0。

花样赛开始了

统计邦德比马场的与众不同之处在于：参加比赛的并不仅有普通马。在接下来的比赛中，3匹斑马将与3匹普通马同场竞技。

在这一轮比赛中，占主导作用的是动物种类，而不是动物本身。也就是说，我们感兴趣的是哪一种动物得到了比赛的哪一种名次。现在请问：按照动物种类进行排名的话，共有几种排列方式？

德比马场设立了特别赌局：只要你押中普通马和斑马的最终排名位置，就给你15：1的赔率。问题是，你应该赌一把吗？

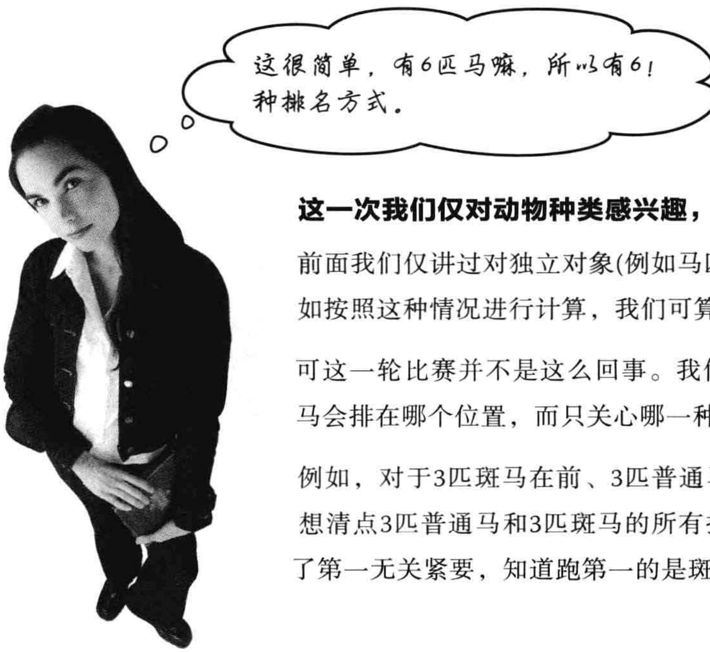


动动脑

你会怎么解答这类问题呢？在以下空白处写下你的想法。

按个体排名与按种类排名不是一回事

如上所述，如果今天的花样赛中有3匹普通马和3匹斑马参赛，如何计算普通马和斑马有几种排名方式?



这一次我们仅对动物种类感兴趣，对动物个体不感兴趣。

前面我们仅讲过对独立对象(例如马匹)进行排名的方式及其数目，假如按照这种情况进行计算，我们可算出正确的结果是6! 。

可这一轮比赛并不是这么回事。我们不再关心哪一匹马或哪一匹斑马会排在哪个位置，而只关心哪一种类的马排在哪个位置。

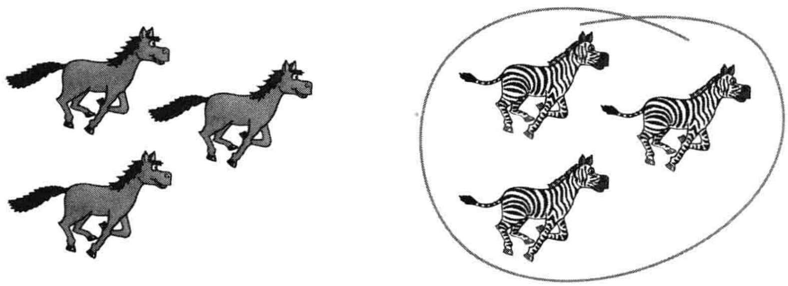
例如，对于3匹斑马在前、3匹普通马在后这种排位情况，我们并不想清点3匹普通马和3匹斑马的所有排名方式。到底是哪一匹斑马跑了第一无关紧要，知道跑第一的是斑马就足够了。



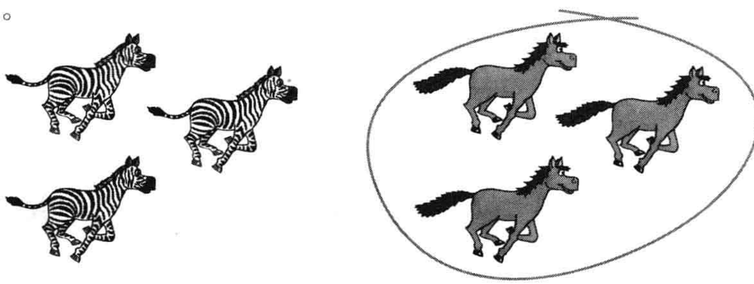
我们需要按种类排列动物

6匹马会有6!种排名方式，但这个答案是假定我们想知道的是单匹普通马(或斑马)的所有可能排名情况。

让我们先看斑马的情况，3匹斑马有3!种排名方式，而上述结果6!中包含这3!种排名情况，但是，由于我们不关心哪一匹斑马排在哪个位置，因此这些排名都是一样的。于是，为了避免重复计算，只需用总数除以3!就行了。



接下来看普通马的情况。3匹普通马有3!种排名方式，而我们先前算出的最终排名结果中包含这3!种排名情况，像斑马的计算方法一样，为了避免重复计算，我们只需用最终结果除以3!就行了。



这意味着按照种类对6匹动物进行排名的数目是：

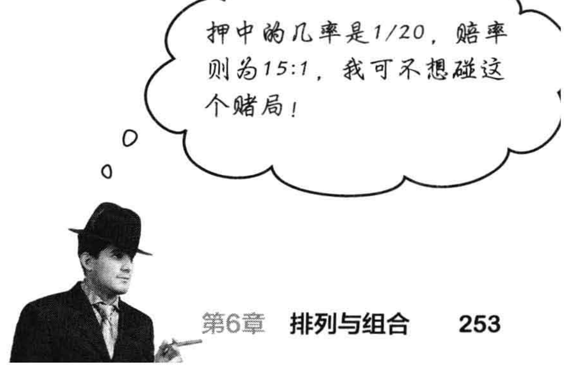
总共有6!种动物排名……

$$\frac{6!}{3!3!} = \frac{720}{6 \times 6} = \frac{720}{36} = 20$$

……但3匹普通马为一类，3匹斑马也为一类，因此用总数除以这些类动物的排名数目。

也就是说，正确押中不同种类动物的排名的概率是1/20。

请翻到下一页，我们将更为详细地讲述这种情况。



推导出用于重复排列的公式

设想你需要清点n个对象的排位方式的总数目，再设想有k个对象是类似对象。

为了求出排位数目，先假定n个对象是独立对象并计算它们的排位数目，用结果除以k种对象（类似对象）的排位方式，得到：

总共有n个对象。

$$\frac{n!}{k!}$$

有k种类似对象。

如果有n个对象，其中k种对象为类似，则排位方式的数目为n!/k!

我们还能进一步推广这个公式。

设想要对n个对象进行排位，其中有一类对象共计k个，另外还有一类对象共计j个，你可以通过下式求出可能的排位方式的数目：

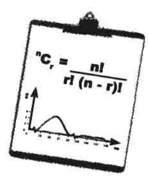
总共有n个对象。

$$\frac{n!}{j!k!}$$

有一类对象有j个，还有另一类对象有k个。

n个对象的排位方式的数目，其中一类有j个类似对象，另一类有k个类似对象。

通常，在计算包括重复对象在内的排位方式数目时，可用总排位方式数目(n!)除以每一类类似对象的排位方式数目(j!, k!等等)。



重要统计量

按类型排位

如果要为n个对象排位，其中包括第一类对象k个，第二类对象j个，第三类对象m个……则排位方式数目的计算式为：

$$\frac{n!}{j!k!m!\ldots}$$



练习

统计邦德比马场决定用自己的比赛进行实验，他们打算办一场有3匹普通马、2匹斑马和5匹骆驼参加的比赛，所有的动物得冠军的可能性都一样。

1. 如果我们对单个动物的情况感兴趣，那么有多少种排名方式？

2. 如果我们只对动物种类的排名感兴趣，那么有几种排名方式？

3. 如果每匹动物赢得冠军的几率一样大，那么5匹骆驼连成一片跑完全程的概率有多大？（假定我们关心的不是单个动物所占据的位置，而是每一类动物所占据的位置。）



练习解答

统计邦德比马场决定用自己的比赛进行实验，他们打算办一场有3匹普通马、2匹斑马和5匹骆驼参加的比赛，所有的动物得冠军的可能性都一样。

1. 如果我们对单个动物的情况感兴趣，那么有多少种排名方式？

有10匹动物，因此有10! = 3,628,800种排名方式。

2. 如果我们只对动物种类的排名感兴趣，那么有几种排名方式？

有普通马3匹，斑马2匹，骆驼5匹。

$$\text{排列数目} = \frac{10!}{3!2!5!}$$

← 有10匹动物。

← 我们将3匹马作为一类，2匹斑马作为一类，5匹骆驼也作为一类。

$$= \frac{3,628,800}{6 \times 2 \times 120}$$

$$= \frac{3,628,800}{1,440}$$

$$= 252$$

3. 如果每匹动物赢得冠军的几率一样大，那么5匹骆驼连成一片跑完全程的概率有多大？（假定我们关心的不是单个动物所占据的位置，而是每一类动物所占据的位置。）

首先，让我们求出5匹骆驼集中在一起跑完全程的方式的数目，为此我们将5匹骆驼划归为一个单一对象，确保它们统一行动。也就是说，如果我们将一群骆驼掺入3匹普通马和2匹斑马中，实际上就需要对6个对象进行排列。

$$\text{排列数目} = \frac{6!}{3!2!}$$

← 1群骆驼＋3匹马＋2匹斑马。

← 我们把3匹普通马当作一类类似对象，把2匹斑马也当作一类类似对象。至于5匹骆驼，则不必除以5!，因为我们把它们计为1个对象了。

$$= \frac{720}{6 \times 2}$$

$$= \frac{720}{12}$$

$$= 60$$

然后，为了求出以上情况的发生概率，我们只需要用骆驼这个整体跑完全程的方式的数目除以所有动物种类跑完全程的全部可能方式的数目，这在上面已经计算过了。

因此，5匹骆驼整体跑完全程的概率为60/252 = 5/21。

世上没有傻问题

问： 在前面的练习中，为什么把5匹骆驼当作一个对象？它们绝对是各自独立的骆驼。

答： 它们的确是**各自独立**的骆驼，但在前面的问题中，我们需要让所有的骆驼成为一个总体，并把这些绑定在一起的骆驼当作一个对象进行处理。

问： 似乎多个不同对象的排位方式的数目与这些对象的分类方式有关。

答： 正确。掌握计算排位方式的方法是一门技术，但还有很大一部分取决于你的思维方式。

关键在于周密地思考实际要解决的问题，还要大量实践。

问： 普通马、斑马和骆驼混在一起比赛的时候多吗？

答： 这是不可能的，不过嘛，这里可是统计邦，统计邦德比马场可以自得其乐嘛。

二十匹马的比赛正在进行

花样赛已经落幕，斑马夺魁。下一轮比赛在20匹马之间进行。



动动脑

该怎么求出20匹马中的前3甲的选取方式呢？

前三甲归属方式有几种？

主赛即将开始，共有20匹马驰骋赛场，我们需要求出前三名的可能排名方式的数目，然后才能算出猜中正确排名的概率。

和前面一样，我们可以先求出马匹占据前3名的方式有几种，然后作出解答。

让我们从第一名开始计算，共有20匹马，即占据第一名的方式有20种，当这个位置被占据后，剩下19匹马占据第二名，再接着就是18匹马占据第三名。



在这场比赛中，我们对剩下的位置被哪匹马占据并不感兴趣，只有前3名才对我们有意义。也就是说，前3名的排列总数是：

$$20 \times 19 \times 18 = 6{,}840$$

于是，准确猜中前三甲正确排名的概率为1/6,840。



我们需要用一个更简炼的方法解决这类问题。

在这里我们只需要将三个数相乘，要是需要将更多的数相乘该怎么办？

我们需要总结出一个公式，以便求出从一个较大的马匹群体中抽出一定数目的马匹进行排名的排名方式总数。

何为排列

讲到这里，我们如何用阶乘重新表示以上算式？

排名方式的数目为20 × 19 × 18，让我们重新推导一下，看看有何结果。

$$20 \times 19 \times 18 = \frac{20 \times 19 \times 18 \times (17 \times 16 \times \ldots \times 3 \times 2 \times 1)}{(17 \times 16 \times \ldots \times 3 \times 2 \times 1)}$$

乘上17!/17!后，式子结果还是一样。

$$= \frac{20!}{17!}$$

这就是同一算式的阶乘表示法。

这是和前面一样的算式，不过现在用阶乘表示。

从20个对象中取出3个对象并进行排位，所得的排位方式的数目有一个正式名称，叫做“排列数目”，如前所述，排列数目的计算方法如下：

$$\frac{20!}{(20-3)!}$$

$$= \frac{2,432,902,008,176,640,000}{355,687,428,096,000}$$

$$= 6,840$$

我们前面得到的也是这个答案。

排列是指从一个较大(n个)对象群体中取出一定数目（r个）对象进行排序，并得出排序方式总数目。

一般说来，从n个对象中取出r个对象的排列数目即n个对象中的每一组对象（r个）的可能排位方式数目，通常写作 ^nP_r，即：

这是对象总数。

$$^nP_r = \frac{n!}{(n-r)!}$$

这是要计算的对象的数目。

所以，若想知道从n个对象中取出r个对象进行排位的排位方式数目，排列算式是个关键。

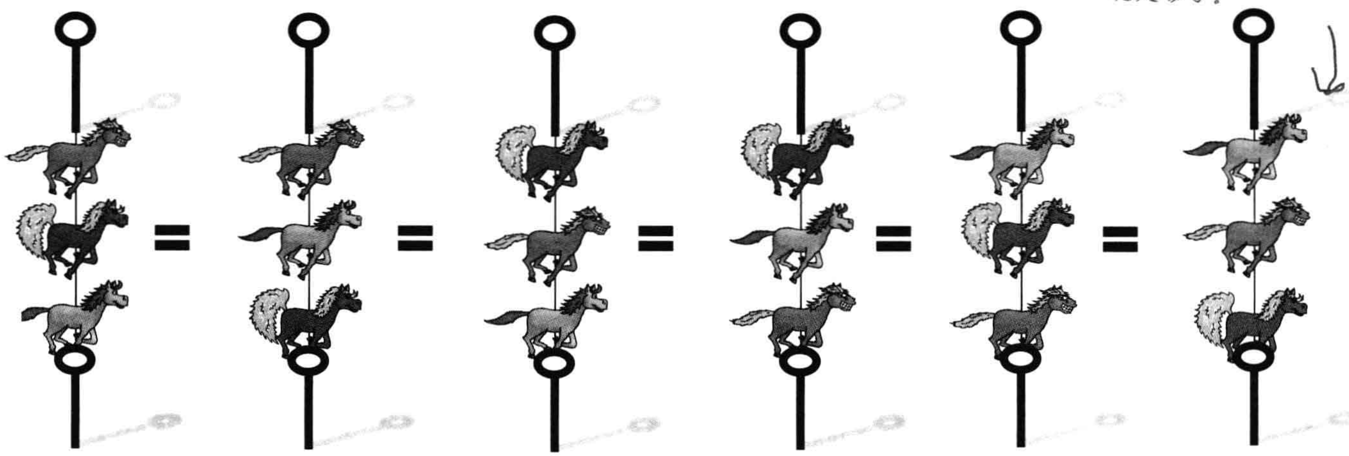


假如马匹排名无关紧要

前面已经讲过从20匹马中取3匹并进行排名的结果，也就是说，我们知道可以给出多少种准确排名。

我们不需要准确知道前3匹马跑完比赛的先后顺序，只需要知道前三名包括哪几匹马就足够了。

而这一次，我们不再想知道排列数目，而想知道前三匹马的**组合**数目——我们仍然需要知道前三名有多少种组合方式，但前三名的确切排名并不细究。



我们该如何解决这类问题呢？

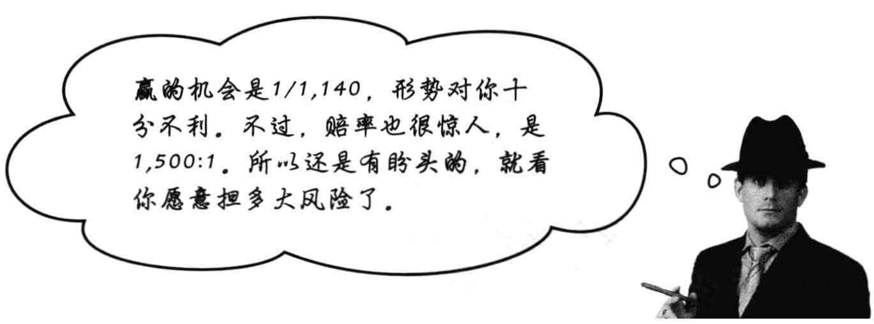
目前，排列数目包括对前3匹马进行确切排名的情况，而3匹马的排名方式有3!种，因此我们用排列数目除以3!，所得结果即为选出占据前三名的马匹但忽略它们的确切排名的选择方式的数目。

结果为：

$$\frac{20!}{3!17!} = \frac{6,840}{3!}$$

$$= 1,140$$

也就是说，选出前3名马匹并进行排名的排列方式有6,840种，但如果不介意排名，则为组合，而组合方式有1,140种。



何为组合

我们前面曾经求出一种计算排列的通用方法，组合其实也有这样一种方法。

一般说来，组合数目即为从n个对象中选取r个对象的选取方式的数目，这时不必知道所选对象的确切顺序。组合数目写作 $^{n}C_{r}$ 即：

$$^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

这是对象的总数目。（指 n）

这是要计算的对象的数目。（指 r）

这一部分的算法与排列的算法相同。（指 n!）

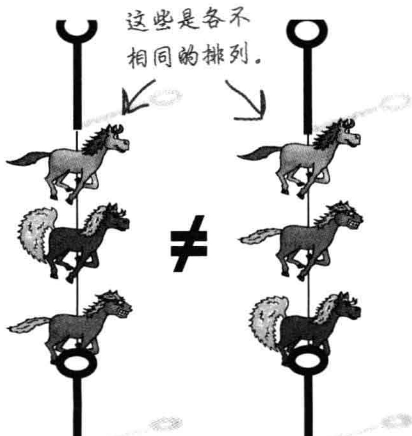
求组合的时候除以一个r!就行了。

那么排列与组合有何区别？

排列

排列是指从一个群体中选取几个对象，在考虑这几个对象的顺序的情况下，求出这几个对象的选取方式的数目。在需要知道每个位置的确切占位情况时，这是一种比组合更明确的方法。

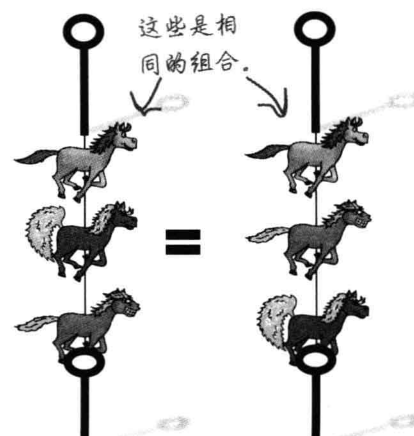
排列：与顺序有关



组合

组合是指从一个群体中选取几个对象，在不考虑这几个对象的顺序的情况下，求出这几个对象的选取方式的数目。在不需要知道每个位置的确切占位情况时，组合是比排列更通用的算法，只要知道所选择的是哪几个对象就足够了。

组合：与顺序无关





组合访谈

本周话题：
顺序重要吗？

Head First：欢迎来到我们的节目，组合先生。

组合：谢谢您的邀请，Head First。

Head First：让我们开门见山吧。很多人都注意到，你和排列十分相似，你对此有同感吗？

组合：我知道人们为什么会这样想，因为我们处理的情况十分相似，我们都关系到从一个群体中选取一定数目的对象。话是这么说，不过我们的相似程度也就仅此而已吧。

Head First：那么你们有什么不同之处呢？

组合：哦，在初学者看来，我们的态度截然不同。排列对顺序很介意，他在选择对象时非常关心选取顺序。他不仅要挑选对象，还要给对象排个位。我是说，他真是的！

Head First：这么说你不这么做？

组合：我绝不！我相信排列的所作所为可谓鞠躬尽瘁，但坦白说，生命苦短，我所关心的是，是否已经从某个群体中选出了对象，若已经选好，那就达到目的了。

Head First：所以你好过排列？

组合：我不愿意说我们两个谁比谁好，好或不好要看具体情况。就拿演奏家打个比方吧……

Head First：演奏家？

组合：是的，很多演奏家都有曲目表，你可以从中选择要演奏家演奏的曲目。

Head First：我想我明白你要说什么了……

组合：这么说，排列和我都对曲目表上的曲目感兴趣，但感兴趣的方式不同。我只要知道曲目表中有哪些曲目就很开心了，而排列却想得更多。他不仅想知道曲目表中的曲目，还想知道曲目的演奏顺序。如果改变曲目顺序，组合不变，但排列就变了。

Head First：谈一谈你的计算方法吧，计算组合的方法和计算排列的方法相似吗？

组合：相似，但略有区别。计算排列的时候，先求n!，接着除以(n－r)!。我的算法很相似，但要再多除以一个r!，通常这会让我变得更小，这是可以理解的，我就是比排列来得痛快。

Head First：通常会变得更小吗？

组合：我换个说法吧，在相似基础数据下，排列永远不会比我小。

Head First：组合先生，谢谢您接受采访。

组合：我很乐意。

世上没有傻问题

问： 我听说过“选取”这样的字眼，这是什么意思？

答： 这是组合的另一个术语。^nC_r的本意是“你有n个对象，选取r个”，因此有时候也称为选择函数。

问： 排列会小于组合吗？

答： 基础数据相同的情况下绝对不会。计算组合要用排列结果再除以一个数值，因此结果肯定变小。

当排列与组合相等时，也就是选取0个或1个对象时，结果最接近你所说的情况。

问： 什么是排列？什么是组合？我又糊涂了。

答： 排列指的是选取对象并关注这些对象的排位顺序，进而得出结果；组合指的是选取对象但不关注这些对象的排位顺序，即可得出结果。

问： 我还是有些糊涂，如果要求从n个对象中选取r个对象的组合，是该写成^nC_r还是^rC_n？

答： 写成^nC_r，记忆窍门：数字越大，位置越高。

问： 它有别的表示方法吗？我想我在什么地方看到过组合的表示方法，不过不是这个样子。

答： 组合的表示方法有不少，我们用的是^nC_r，但另外还有一种表示法，即：

$$\binom{n}{r}$$

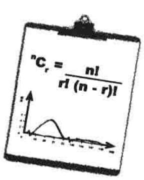
问： 排列和组合是否的确十分重要？

答： 没错，尤其是组合，本书后文还多有涉及。需要的时候要记得用哦。

问： 计算排列和组合的情况似乎和计算类似对象的情况很相似，对吗？

答： 过程相似。在计算类似对象时，是用排位方式的总数目除以类似对象的分类数目。

对于排列可以这样理解：你选取互不同类的所有对象进行计算，因此用n!除以(n-r)!；对于组合可以这样理解：你选取的对象都是同类，所以要用排列的数目再除以r!。



重要统计量

排列

如果从n个对象中选取r个对象，则排列数目为：

$$^nP_r = \frac{n!}{(n-r)!}$$

组合

如果从n个对象中选取r个对象，则组合数目为：

$$^nC_r = \frac{n!}{r!(n-r)!}$$



练习

数据邦全明星篮球队即将参加一场比赛，在册队员12名，同一时间允许5名队员上场比赛。

1. 同一时间上场比赛的队员有几种出场方式?

2. 教练指定了3名队员做投篮主力。如果这3名主力是随机选择的，那么3名主力在同一时间上场的概率有多大?

练习

现在该算扑克牌概率了，看看你怎么应付。

一副牌有52张，一手牌有5张，拿一手牌的方式有几种？

全部同花的10、J、Q、K、A组成一个同花大顺。拿到这种扑克牌组合的概率是多少？用上一题的答案帮忙解答。

四张数字相同的牌组成一个“炸弹”，再加一张牌就成一手。拿到这种扑克牌组合的概率是多少？

五张花色相同的牌组成一手同花牌。拿同花牌的概率是多少？



数据邦全明星篮球队即将参加一场比赛，在册队员12名，在同一时间允许5名队员上场比赛。

1. 同一时间上场比赛的队员有几种出场方式?

在册队员有12名，我们需要计算从其中挑选5名队员的挑选方式的数目。不需要对挑选出来的队员进行排序，因此可以用组合进行计算。

$$^{12}C_5 = \frac{12!}{5!(12-5)!}$$

$$= \frac{12!}{5!7!}$$

$$= 792$$

2. 教练指定了3名队员做投篮主力。如果这3名主力是随机选择的，那么3名主力在同一时间上场的概率有多大?

让我们先算3名主力同时上场的方式的数目。

如果3名主力同时上场，就表示还剩下2个位置供其他队员填补。我们需要求出从剩余9名队员中选取2名队员填补上述2个位置的组合数目。

$$^{9}C_2 = \frac{9!}{2!(9-2)!}$$

$$= \frac{9!}{2!7!}$$

$$= 36$$

这就是说，3名主力同时上场的概率为：

$36/792 = 1/22$

练习解答

现在该算扑克牌概率了，看看你怎么应付。

一副牌有52张，一手牌有5张，拿一手牌的方式有几种？

一副牌有52张，我们需要从中选择5张。

$$ {}^{52}C_5 = \frac{52!}{47!5!} = 2,598,960 $$

全部同花的10、J、Q、K、A组成一个同花大顺。拿到这种扑克牌组合的概率是多少？用上一题的答案帮忙解答。

每一种花色出现这种组合的情况有1种，总共4种花色。也就是说，拿到同花大顺的方式有4种。

$$ P(\text{同花大顺}) = \frac{4}{2,598,960} $$
$$ = 1/649,740 $$
$$ = 0.0000015 $$

四张数字相同的牌组成一个“炸弹”，再加一张牌就成一手。拿到这种扑克牌组合的概率是多少？

让我们从“炸弹”着手，总共有13种可能，即组成“炸弹”的方式有13种，只要选出一副“炸弹”，就剩下48张牌。也就是说，这样一手牌的组成方式的数目为：$13 \times 48 = 624$。

$$ P(\text{炸弹}) = \frac{624}{2,598,960} $$
$$ = 1/4165 $$
$$ = 0.00024 $$

五张花色相同的牌组成一手同花牌。拿同花牌的概率是多少？

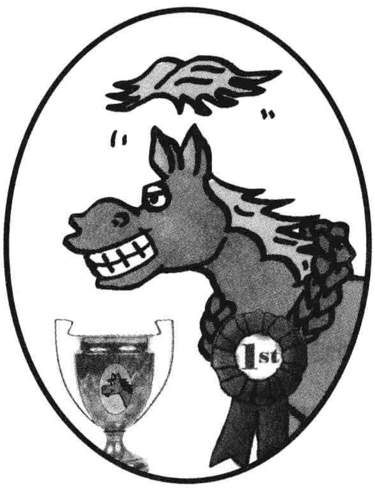
为了求出可能的组合的数目，先求一套同花牌选取方式的数目，这个数目为4，然后选取这套花色中的5张牌。每种花色有13张牌，于是所求组合数目为：

$$ 4 \times {}^{13}C_5 = \frac{4 \times 13!}{8!5!} $$
$$ = 4 \times 1287 = 5148 $$

$$ P(\text{同花}) = \frac{5148}{2,598,960} $$
$$ = 33/16660 $$
$$ = 0.00198 $$

比赛结束

二十匹马的比赛已经结束，冠军拉托，翠香屈居第二，福福季军。如果你当初决定押这三匹马赢，那你就发了！



统计邦德比马场
本年度冠军：
拉托



第2名：
翠香



第3名：
福福

在本章中，你学习了如何处理各种排名、排位和排列，以及如何在不一一列举各种可能性的情况下快速算出可能的排列、组合的数目。

这些知识将大大提高你求概率和作统计的能力。请接着读下去，我们会继续指点你练就更强功力。

7 几何分布、二项分布及泊松分布

坚持离散



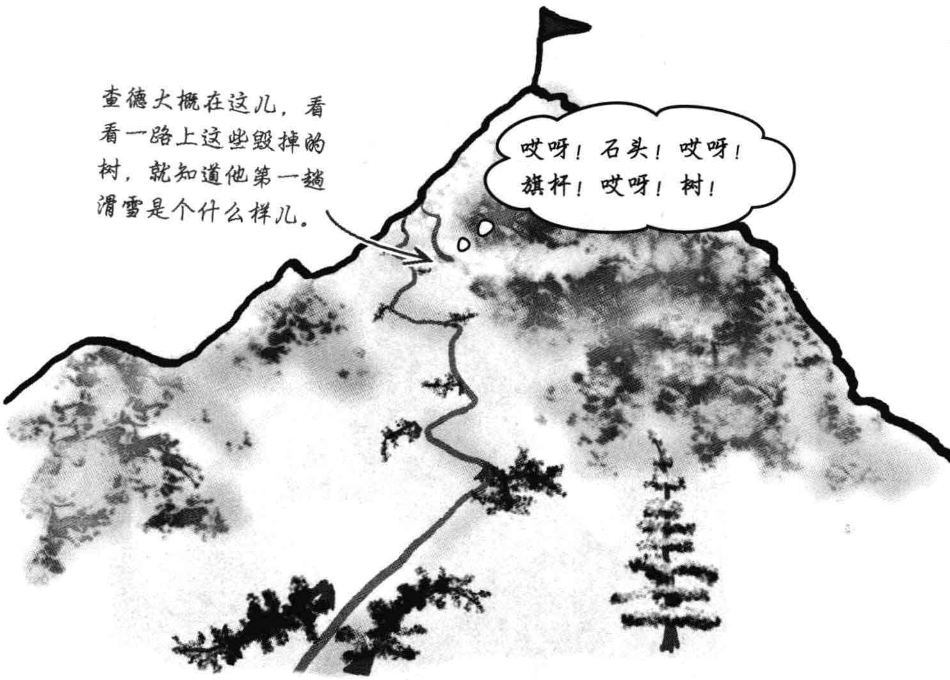
计算概率分布颇为费时。

前面讲到如何计算和利用概率分布，不过，如果**方法更简单一些，计算速度更快一些**，效果岂不更好？在本章中，我们将介绍一些**特殊的概率分布**，这些概率分布有着十分固定的模式。只要懂得这些模式并善加利用，就能**以前所未有的速度计算概率、期望、方差**。接着读吧，让我们一起来认识几何分布、二项分布及泊松分布。

倒霉的滑雪者查德



查德喜欢滑雪，但他是个事故大王，哪怕雪坡上只有孤零零的一棵树，他也准能撞上去。查德希望自己不要总是撞在树上，滚在雪里，他的保险费如今可是一笔大开销。



查德对自己在雪坡上的表现寄望甚高：他的自尊，他对雪上美女的成功追逐，他的保险，为此他愿意冒丢人现眼、断手断脚、保险大打折扣的风险学习新的滑雪技巧，但必须保证他试滑不到10次就能获得成功。

查德不出事故顺利滑至坡底的概率是0.2，他打算不停尝试，直至大功告成。在取得第一次成功后，他将停止滑雪，高唱凯歌回小旅馆。

动动笔

现在来练习练习你求概率的技术。查德在任意一次试滑中(假定每一次试滑都是独立的)不出事故顺利抵达坡底的概率均为0.2。如果需要试滑两次，概率如何？他试滑一次或两次就能成功滑至坡底的概率是多大？记住：当他获得首次成功后，就打算歇手不干。

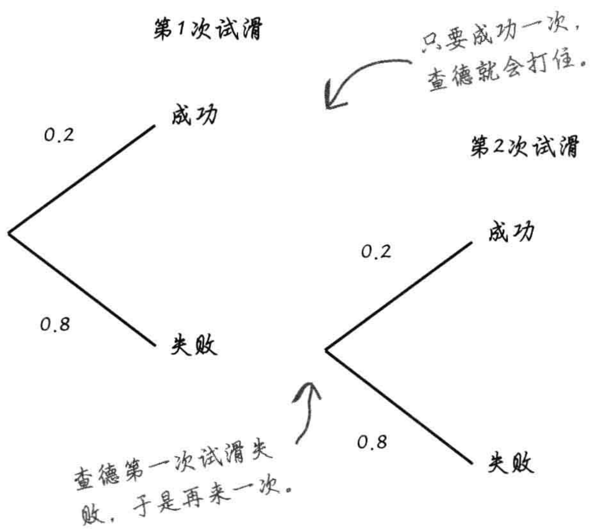
查德可谓百折不挠，在任何一次滑雪中遭遇的跌倒摔跤都不会影响他下一次的表现。

提示：你可能打算画一棵概率树，以便让问题直观可视。

动动笔解答

现在来练习练习你求概率的技术。查德在任意一次试滑中(假定每一次试滑都是独立的)不出事故顺利抵达坡底的概率均为0.2。如果需要试滑两次，概率如何？他试滑一次或两次就能成功滑至坡底的概率是多大？记住：当他获得首次成功后，就打算歇手不干。

下面是一棵概率树，其中给出了前两次试滑的概率，有了这些就能算出概率了。



如果用X表示最终滑到坡底需要试滑的次数，则：

$P(X = 1) = P(\text{第1次试滑成功})$

$= 0.2$

$P(X = 2) = P(\text{第2次试滑成功} \cap \text{第1次试滑失败})$

$= 0.2 \times 0.8$

$= 0.16$

$P(X \leq 2) = P(X = 1) + P(X = 2)$ ← 由于这些概率相互独立，因此可以相加。

$= 0.2 + 0.16$

$= 0.36$

我们需要求出查德的概率分布

现在，你已经求出了查德在雪坡上试滑不出3次就能成功的概率，不过，如果你需要了解他试滑不出10次(因为保险的原因)就成功的概率，那该怎么办？20次呢？100次呢？

相对于每一次都老老实实地从头开始计算概率，概率分布可能更方便。为此，我们需要指出查德最终到达坡底需试滑次数的每一种可能性，并算出相应概率。



这样做有问题，因为可能次数无穷无尽。

只要尚未试滑成功，查德就会不停地试下去。他可能要试1次，10次，100次……甚至1,000次。查德到底什么时候会获得首次成功？谁也不能确定。



即使可能次数无穷无尽，还是有办法求出它的概率分布的。

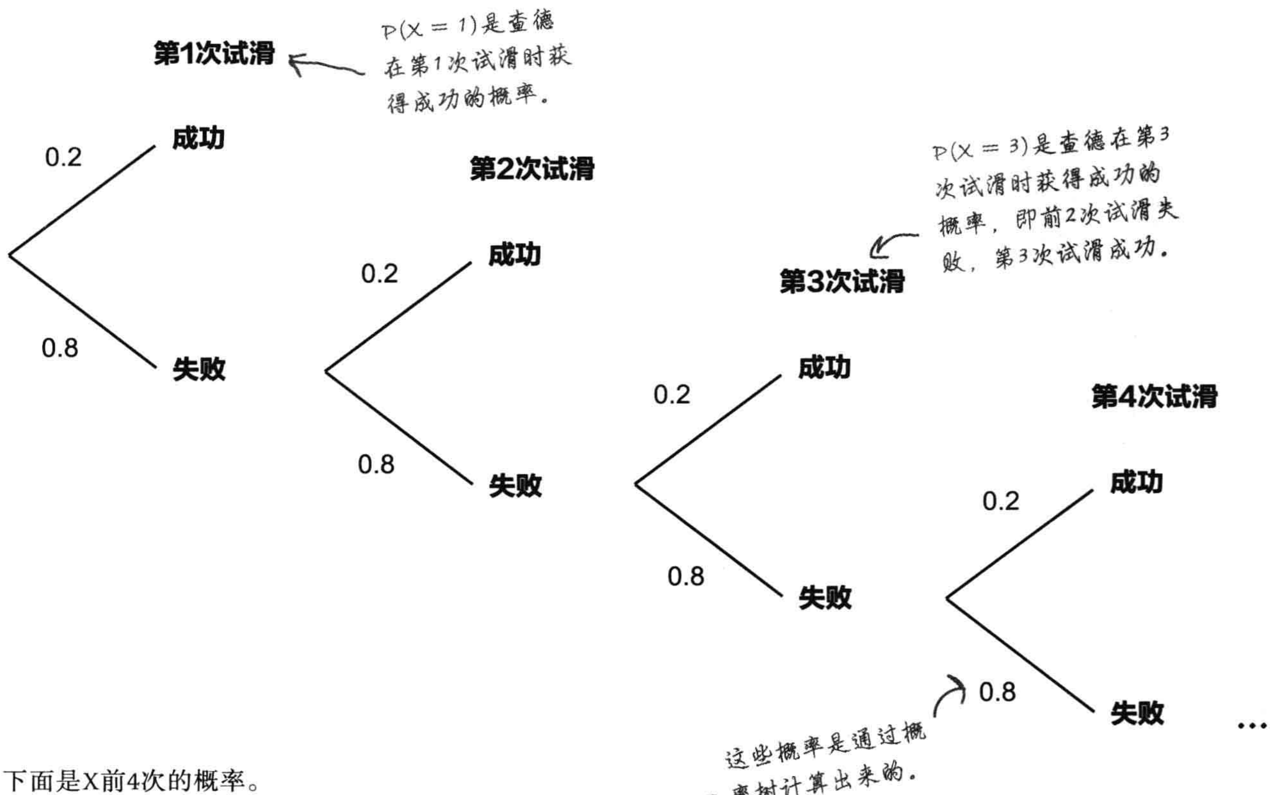
这其实是一种特殊的概率分布，这种概率分布具有一些特殊属性，能够简化概率、数学期望，以及方差的计算。

让我们看看如何处理。

这种概率分布有一种固定模式

让我们用变量X表示查德为了在雪坡上取得一次成功而需要经历的试滑次数。查德只需要成功一次即可，此后他将停止试滑。

让我们先看前4次试滑，据此计算X的前4个数值的概率。然后，我们可以看看是否存在某种固定模式能帮助我们轻松地算出其余数值的概率。



下面是X前4次的概率。

x	P(X = x)
1	0.2
2	0.8 × 0.2 = 0.16
3	0.8 × 0.8 × 0.2 = 0.128
4	0.8 × 0.8 × 0.8 × 0.2 = 0.1024

注意，每个概率都是0.8的幂与0.2的幂的乘积。



练习

下表用于填写X取不同数值时的相应概率，请填写表格，写出试滑次数为x时的概率，并指出每种情况下，0.8的幂和0.2的幂(0.8和0.2在P(X = x)中出现的次数)分别是多少。

x	**P(X = x)**	**0.8的幂**	**0.2的幂**
1	0.2	0	1
2	0.8 × 0.2	1	1
3	$0.8^2 \times 0.2$	2	
4			
5			
r			

r是x的一个特定值，但现在还不告诉你到底是哪个值。你能猜一猜r的相应概率是多少吗？

这一大块空白是留给你做计算的。



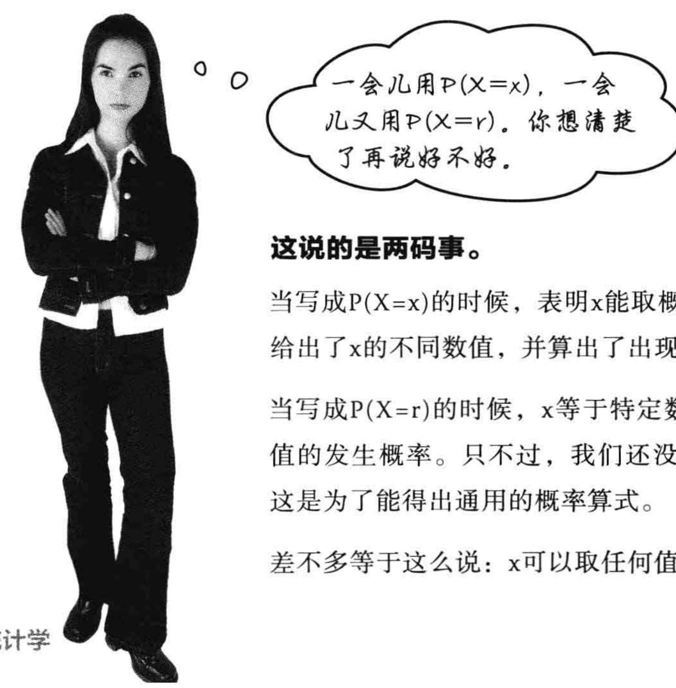
下表用于填写X取不同数值时的相应概率，请填写表格，写出试滑次数为x时的概率，并指出每种情况下，0.8的幂和0.2的幂(0.8和0.2在P(X = x)中出现的次数)分别是多少。

x	**P(X = x)**	**0.8的幂**	**0.2的幂**
1	0.2	0	1
2	0.8×0.2	1	1
3	$0.8^2 \times 0.2$	2	1
4	$0.8^3 \times 0.2$	3	1
5	$0.8^4 \times 0.2$	4	1
r	$0.8^{r-1} \times 0.2$	$r - 1$	1

当X＝4时，查德先失败3次，第4次成功。
由于单次试滑的成功概率为0.2，失败概率为0.8，因此$P(X = 4)$为$0.8 \times 0.8 \times 0.8 \times 0.2$。

当X＝5时，查德先失败4次，第5次成功。即
$P(X = 5) = 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.2$.

那么，$P(X = r)$是多少呢？若查德在第r次试滑时成功，则肯定已经先失败过$(r - 1)$次，于是
$P(X = r) = 0.8 \times 0.8 \times \cdots \times 0.8 \times 0.2$，即表达式中的0.8取$(r - 1)$次幂。



这说的是两码事。

当写成P(X=x)的时候，表明x能取概率分布中的任何值。我们在上表中给出了x的不同数值，并算出了出现每种数值的概率。

当写成P(X=r)的时候，x等于特定数值r，我们要求的就是这个特定数值的发生概率。只不过，我们还没有指定这个特定数值r到底是多少，这是为了能得出通用的概率算式。

差不多等于这么说：x可以取任何值，包括固定数值r。

概率分布可以用代数式表示

如你所见，查德的滑雪试验有其特定模式。每一个概率都是0.8和0.2的乘积，利用下式，你能迅速算出任意次数r的概率：

$$P(X = r) = 0.8^{r-1} \times 0.2$$

即，如果要求P(X = 100)，你不需要画出一棵硕大无比的概率树，也不用把每一次试滑的情形想得清清楚楚，只要这样算就行：

$$P(X = 100) = 0.8^{99} \times 0.2$$

我们可以进一步总结这个公式。如果用p代表单次试滑的成功概率，则失败的概率为$1-p$，我们将此概率称为q，于是可以用下式计算任何具有这一性质的概率：

$$\mathbf{P(X = r) = q^{r-1}\,p}$$

(r－1)次失败，1次成功。在我们的例子中，p＝0.2，q＝0.8。

真是没用，
老是失败<抽泣>

q

q等于1-p。如果p代表成功概率，则q代表失败概率。

这个公式叫做概率的**几何分布**。

世上没有傻问题

问： 总结这个公式有什么意义呢？这只是我们所计算的一个特别问题而已。

答： 总结这个公式是为了用到其他类似问题上。如果我们能够总结出这类问题的结果，以后碰到类似情况时就能加快计算速度。

问： 你说过，我们需要求出P(X=r)的表达式，r是什么？

答： P(X=r)表示“X等于数值r的概率”，其中r是为了取得首次成功所需进行的试验次数。例如，如果想求出P(X=20)，那么就可以用20代替r，这样就能迅速求出概率。

问： 为什么用字母r呢？为什么不用其他字母呢？

答： 使用字母r便于将结果推广至任何特定数值，其实我们也可以用其他字母，不过常用的就是r。

问： 如果可能出现的结果无穷无尽，我们如何求出概率分布？

答： 我们不用为了得出概率分布而一一列出每一种可能结果，关键在于通过某种方式描述每一种可能结果，概率计算公式就是这样一种方式。

问： 查德的滑雪技巧难道自始至终都不会提高吗？说每一次试滑的成功概率都是0.2，这现实吗？

答： 你的想法有道理。不过在滑雪问题上，查德实在是非常倒霉，我们不得不假定他的技巧没有提高——也就是说，他滑雪成功的概率符合几何分布。

几何分布细细看

我们说过，查德的滑雪壮举是**几何分布**的一个实例。几何分布包含以下条件：

1. 进行一系列相互独立的试验。
2. 每一次试验都既有成功的可能，也有失败的可能，且单次试验的成功概率相同。
3. 你主要感兴趣的是，为了取得第一次成功需要进行多少次试验。

如果你所碰到求概率的情况满足这几个条件，那么就可以用几何分布的公式帮助你速战速决。这里有一个重要提示：我们用了“成功”这个词表示我们感兴趣的事件成为事实，假如我们希望看到的事件具有负面含义，从统计学的角度看，这个负面事件仍然可算得是一个“成功”事件。

让我们用变量X表示**为了取得第一次成功所需进行的试验次数**，即，为了让我们感兴趣的事件发生而需要进行的试验次数。

为了求出X取特定数值r的概率，可以用下式进行快速计算：

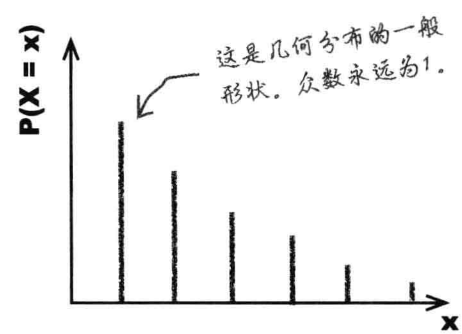
$$P(X = r) = p\,q^{r-1}$$

其中p为成功概率，q=1-p为失败概率。即，为了在第r次试验时取得成功，首先要失败(r-1)次。

几何分布的形状十分独特。

当r=1时，P(X=r)达到最大值，随着r增大，P(X=r)逐渐下降。注意，取得成功的概率在第一次试验时最大，也就是说，**任何几何分布的众数都永远是1**，因为1是具有最大概率的数。

虽然看似有违直觉，但是，可能性最大的情况却是：仅需尝试一次即可成功。



几何分布对不等式同样有用

像求解几何分布的准确概率一样，对于涉及不等式的概率，也有一种简便的求解方法。

让我们从P(X > r)讲起。

P(X > r)指的是为了取得第一次成功需要试验r次以上的概率。为了让需要进行的试验次数大于r，意味着前r次试验必须以失败告终。也就是说，将失败概率乘上r次就是所求的概率。

$$P(X > r) = q^r$$

为了让取得成功时的试验次数大于r，必须先有r次失败。

这个公式中不需要出现p，因为我们不需要确切地知道哪一次试验是成功的，只要知道试验次数必须大于r就行了。

我们可以利用这个公式求出P(X ≤ r)，即为了取得一次成功而需要尝试r次或r次的以下概率。

如果将P(X ≤ r)和 P(X > r)相加，结果必为1，即：

$$P(X \le r) + P(X > r) = 1$$

或

$$P(X \le r) = 1 - P(X > r)$$

这是因为P(X ≤ r)与P(X > r)是两种对立的情况，P(X ≤ r) = 1 - P(X > r)。

由此得出：

$$P(X \le r) = 1 - q^r$$

从上式可知，$P(X > r) = q^r$

于是我们用q^r代替P(X > r)，得出这个公式。

如果一个变量X的概率符合几何分布，且单次试验的成功概率为p，则可以写作：

$$X \sim Geo(p)$$

这个简明表达式的意思是“X符合几何分布，其中成功概率为p”。



几何分布的期望模式

前面已经求出查德为了成功滑到坡底而需要试滑的次数，但如果想求期望和方差呢？知道期望用处很多，例如，在数学期望已知的情况下，就可以得出查德在成功之前试滑次数的期望值。

提示一下，期望就是你期望得到的平均值，有点儿像均值，不过是概率分布的均值。

方差则是对偏差的量度。

还记得本书前面部分是如何求期望的吗？E(X)可以通过$\sum xP(X=x)$进行计算。这个例子有无穷多个概率。不过，我们可以先算算前面几个数值，看看是否存在某种固定模式。

下面是x的前几个数值，其中X ~ Geo(0.2)

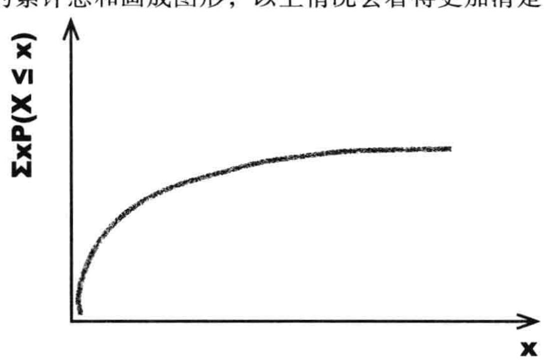
这是xP(X = x)的累计总和。

x	P(X = x)	xP(X = x)	xP(X ≤ x)
1	0.2	0.2	0.2
2	0.8 × 0.2 = 0.16	0.32	0.52
3	0.8^2 × 0.2 = 0.128	0.384	0.904
4	0.8^3 × 0.2 = 0.1024	0.4096	1.3136
5	0.8^4 × 0.2 = 0.08192	0.4096	1.7232
6	0.8^5 × 0.2 = 0.065536	0.393216	2.116416
7	0.8^6 × 0.2 = 0.0524288	0.3670016	2.4834176
8	0.8^7 × 0.2 = 0.04194304	0.33554432	2.81894608

能看出xP(X = x)的特点吗？

xP(X=x)的数值一开始很小，接着越变越大，直到x=5。当x大于5时，数值又开始减小，并且随着x的变大而继续减小。X越来越大，xP(X=x)越来越小，直到几乎不能使累计总和发生变化。

如果将xP(X = x) 的累计总和画成图形，以上情况会看得更加清楚：



期望是1/p

将$xP(X = x)$ 的累计总和画成图形后,可以看出，随着x变大，累计总和越来越接近一个特定数值：5。实际上，经过无穷多次试验后，$xP(X = x)$的累计总计正是等于5。即：

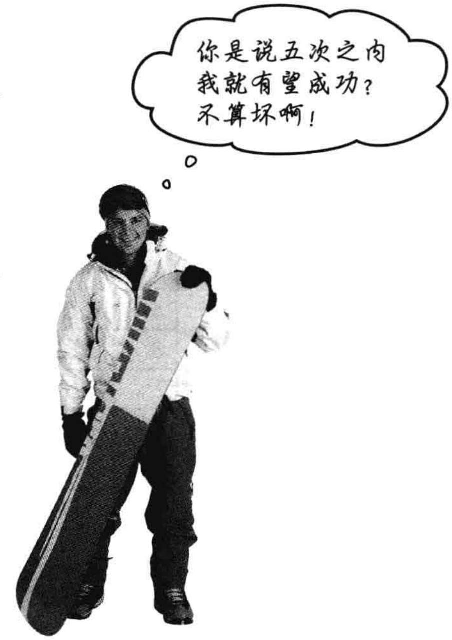
$$E(X) = 5$$

上式的意义很直观：单次试验的成功概率为0.2，可以理解为5次尝试中有一次尝试趋向于成功，因此我们可以期望查德尝试5次即获成功。

以上情况可以推而广之至任意数值p。如果$X \sim Geo(p)$，则：

$$E(X) = \frac{1}{p}$$

期望等于1除以成功概率



我们不仅能求出几何分布的期望，还能求出方差。

动动笔

让我们看看是不是能用求期望的同样方式求出几何分布的方差表达式。填写下表，有何发现？

x	$P(X = x)$	$x^2P(X = x)$	$x^2P(X \leq x)$
1	0.2		
2	$0.8 \times 0.2 = 0.16$		
3	$0.8^2 \times 0.2 = 0.128$		
4	$0.8^3 \times 0.2 = 0.1024$		
5	$0.8^4 \times 0.2 = 0.08192$		
6	$0.8^5 \times 0.2 = 0.065536$		
7	$0.8^6 \times 0.2 = 0.0524288$		
8	$0.8^7 \times 0.2 = 0.04194304$		
9	$0.8^8 \times 0.2 = 0.033554432$		
10	$0.8^9 \times 0.2 = 0.0268435456$		

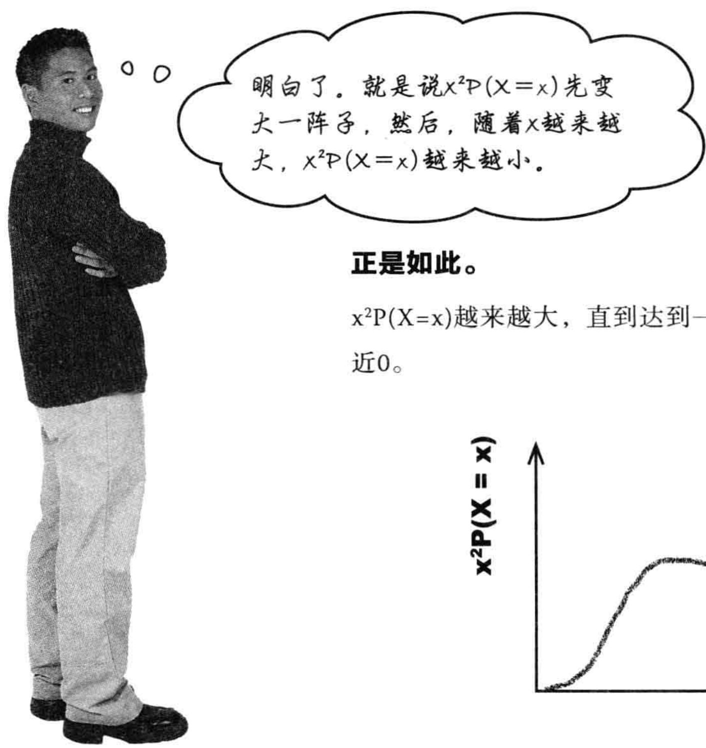
记住：方差的计算方法是$E(X^2) - E^2(X)$.

动动笔解答

让我们看看是不是能用求期望的同样方式求出几何分布的方差表达式。填写下表，有何发现？

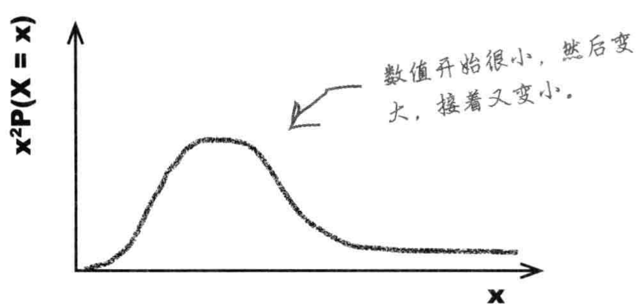
x	P(X = x)	x^2P(X = x)	x^2P(X ≤ x)
1	0.2	0.2	0.2
2	0.8 × 0.2 = 0.16	0.64	0.84
3	0.8^2 × 0.2 = 0.128	1.152	1.992
4	0.8^3 × 0.2 = 0.1024	1.6384	3.6304
5	0.8^4 × 0.2 = 0.08192	2.048	5.6784
6	0.8^5 × 0.2 = 0.065536	2.359296	8.037696
7	0.8^6 × 0.2 = 0.0524288	2.5690112	10.6067072
8	0.8^7 × 0.2 = 0.04194304	2.68435456	13.29106176
9	0.8^8 × 0.2 = 0.033554432	2.717908992	16.00897075
10	0.8^9 × 0.2 = 0.0268435456	2.68435456	18.69332531

这一次，x^2P(X = x) 不断增加，直到x达到10。当x达到10之后，x^2P(X = x)再次开始下降。



正是如此。

x^2P(X=x)越来越大，直到达到一个特定值，然后又开始减小，最终变得非常接近0。



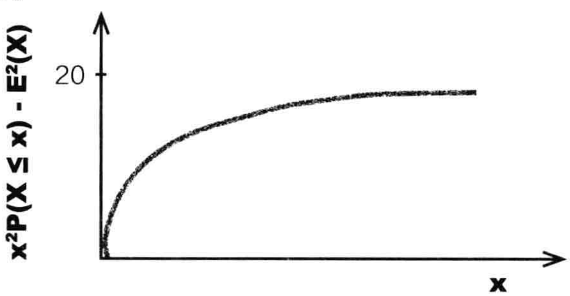
求当前分布的方差

以上分析如何帮助我们求出查德成功滑到坡底需要试滑的次数的方差？

通过下式可算出概率分布的方差：

$$Var(X) = E(X^2) - E^2(X)$$

即，算出 $\sum x^2P(X = x)$，减掉E(X)的平方，以x为横轴画出所得结果的图形，这时可以看出Var(X)的模式是：随着x上升而上升。下面是 $x^2P(X \leq x) - E^2(X)$的图形。



随着x变大，$x^2P(X \leq x) - E^2(X)$越来越接近一个特定数值，这里是20。

和讨论数学期望的时候一样，方差的规律归结如下。如果$X \sim Geo(p)$，则

$$\mathbf{Var(X) = \frac{q}{p^2}}$$



几何分布简明指南

下面是有关几何分布的简明总结，你可能用得上：

何时使用几何分布？

进行多次相互独立的试验时可使用几何分布，每一次试验都存在成功或失败的可能，而你感兴趣的是为了取得第一次成功需要试验多少次。

如何计算概率？

可使用以下方便易用的公式。P为单次试验的成功概率，q=1-p，X是为了取得第一次成功而需要试验的次数，这时我们说X ~ Geo(p)。

$$P(X = r) = p\,q^{r-1}$$

在第r次试验时取得第一次成功的概率。

$$P(X > r) = q^r$$

需要试验r次以上才能取得第一次成功的概率。

$$P(X \le r) = 1 - q^r$$

需要试验r次或不到r次即可取得第一次成功的概率。

如何计算方差和期望？

公式如下：

$$E(X) = 1/p \qquad Var(X) = q/p^2$$

世上没有傻问题

问： 这些公式可靠吗？任何时候都能用来求概率和期望吗？

答： 只要是几何分布，就可以用这些速算公式，因为这些公式正是针对几何分布的简便算法。如果所处理的问题不符合几何分布模型，那么不要用这些简便算法。

别忘了，几何分布的应用条件是：进行多次相互独立的试验（因此每次试验的概率保持不变），每一次试验都存在失败或成功的可能性，而你感兴趣的是：为了取得第一次成功需要进行多少次试验。

问： 如果是其他情况呢？例如试验次数一定，要求成功次数呢？

答： 不能使用几何分布，你说的情况不符合几何分布的模型。不过别担心，会有其他方法的。

问： 我要把这些速算法都学会吗？

答： 如果你要处理几何分布问题，知道这些公式会大大节省你的时间；如果你是为了参加统计学考试，那么看看考试大纲是否要求学会这些内容。

问： 为什么几何分布用到p和q？

答： p代表英文单词“probability”，即“概率”，在几何分布中，代表的是单次试验的成功概率。q在统计学中往往代表1-p，也就是p'。本章以及本书后文将会大量出现这些字母。

化身滑雪者



另一位滑雪者不出意外顺利滑至坡底的概率是0.4。你的任务是假装自己是这位滑雪者，算出以下情况下的概率。

1. 第一次滑雪失败，第二次滑雪成功的概率。

2. 第4次或不足4次就滑雪成功的概率。

3. 4次以上才能获得成功的概率。

4. 你所期望的为了获得成功而需要试滑的次数。

5. 试滑次数的方差。

化身滑雪者解答



另一位滑雪者不出意外顺利滑至坡底的概率是0.4。你的任务是假装自己是这位滑雪者，算出以下情况下的概率。

让我们使用$X \sim Geo(0.4)$进行解答，其中X为这位滑雪者为了顺利滑至坡底而需要经历的试滑次数。

1. 第一次滑雪失败，第二次滑雪成功的概率。

$$P(X=2) = p \times q$$
$$= 0.4 \times 0.6$$
$$= 0.24$$

2. 第4次或不足4次就滑雪成功的概率。

$$P(X \leq 4) = 1 - q^4$$
$$= 1 - 0.6^4$$
$$= 1 - 0.1296$$
$$= 0.8704$$

3. 需要滑雪4次以上才能获得成功的概率。

$$P(X > 4) = q^4$$
$$= 0.6^4$$
$$= 0.1296$$

或者可以这样求：

$$P(X > 4) = 1 - P(X \leq 4)$$
$$= 1 - 0.8704 = 0.1296$$

4. 你所期望的为了获得成功而需要试滑的次数。

$$E(X) = 1/p$$
$$= 1/0.4$$
$$= 2.5$$

5. 试滑次数的方差。

$$Var(X) = q/p^2$$
$$= 0.6/0.4^2$$
$$= 0.6/0.16$$
$$= 3.75$$



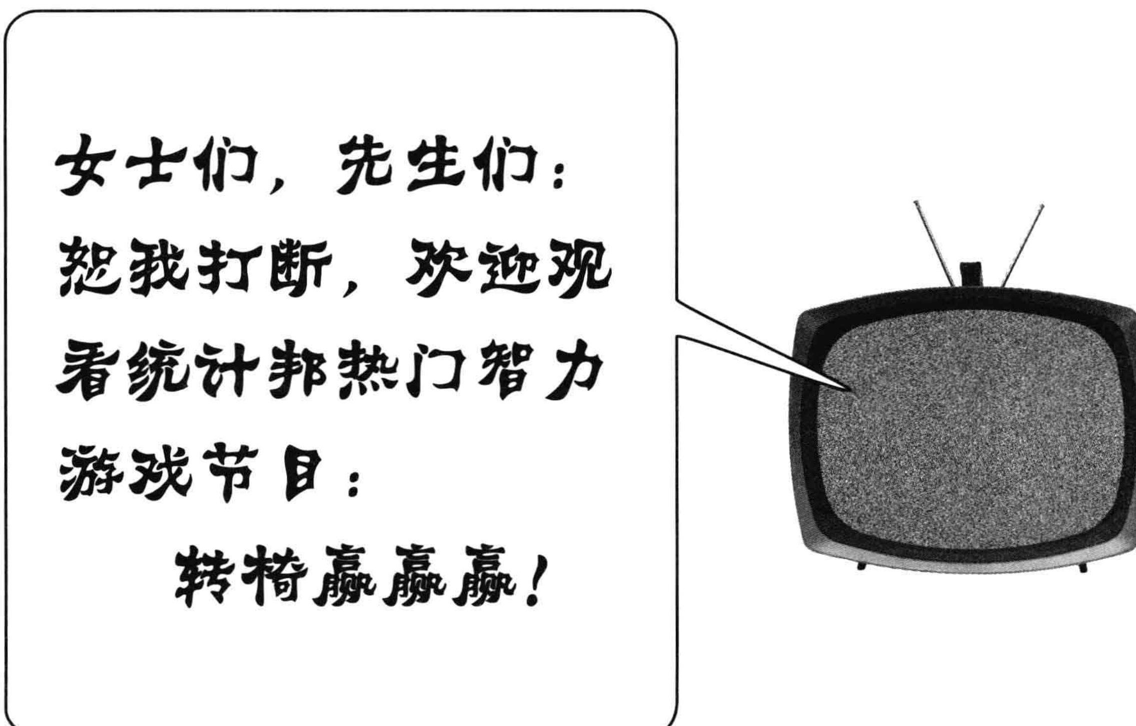


你已经掌握了几何分布

多亏你懂得几何分布这门技术，查德不仅知道自己在试滑多少次之后可能成功滑到坡底的概率，还能知道他能够期望自己滑多少次就获得成功，以及存在多大变数。

只要试滑5次就能成功滑至坡底，方差为20，这样的期望让他倍增自信 —— 他不用伤痕累累就能让那些美女刮目相看了。

继续前进……




WHO WANTS TO WIN A
SWIVEL CHAIR
WHO
WIN A
大家好，欢迎观看统计邦热门节目“转椅赢赢赢”。今晚我们准备了一些搞怪难题，希望您福星高照。

今天我们可为您准备了不少难题，现在开始！第一轮3个问题，每个问题有4个备选答案，您可以现在就带着鼓励奖离场，也可以选择继续，击败对手进入下一轮后，您就离转椅近了一步。第一轮："关于我"。祝您好运！



动动笔

下面是第一轮的提问，都是关于游戏主持人的问题。请在正确答案旁边打勾。

1. 他喜欢什么颜色？

- **A:** 红色
- **B:** 蓝色
- **C:** 绿色
- **D:** 黄色

2. 他的生日在几月份？

- **A:** 1月
- **B:** 2月
- **C:** 3月
- **D:**4月

3. 人们最喜欢他哪一点？

- **A:** 长相好看
- **B:** 有魅力
- **C:** 有幽默感
- **D:** 机智

世上没有傻问题

问： 讲到一半来个智力游戏干什么？还是接着讨论概率分布吧。

答： 还是在讨论着。智力游戏是另一种概率分布的理想案例，读下去你就会明白的。

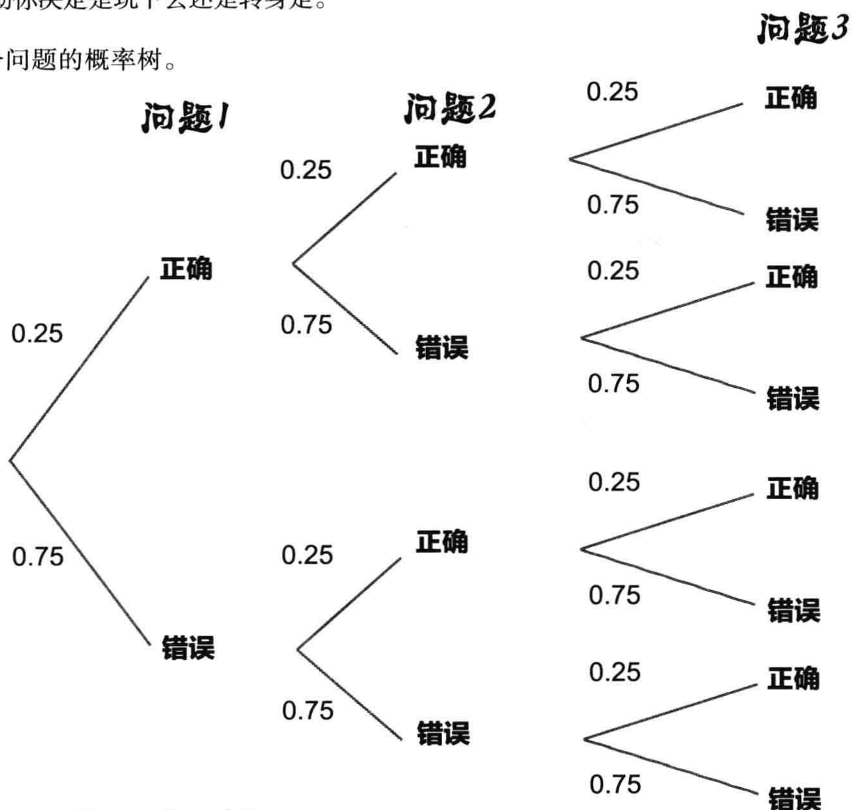
问： 我不知道这些问题的答案，怎么办？

答： 不知道答案可以随机答嘛，好好猜吧，有可能得大奖呢。

玩下去，还是转身走？

你不太可能对游戏主持人那么了解，所以这些问题应该是答不上来的。所以，让我们看看，如果随机回答问题，是否能求出答对的题数的概率分布，这会帮助你决定是玩下去还是转身走。

这是3个问题的概率树。



动动笔

这类问题的概率是多大？可以看出什么规律？用X代表答对的题数，共3题。

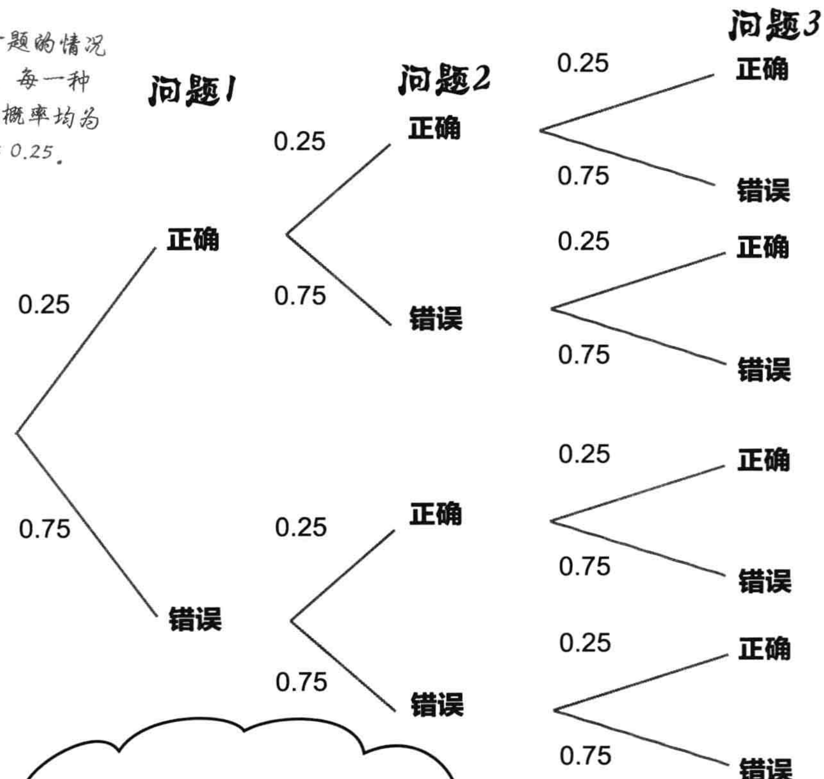
x	P(X = x)	0.75的幂	0.25的幂
0	0.75^3	3	0
1			
2			
3			

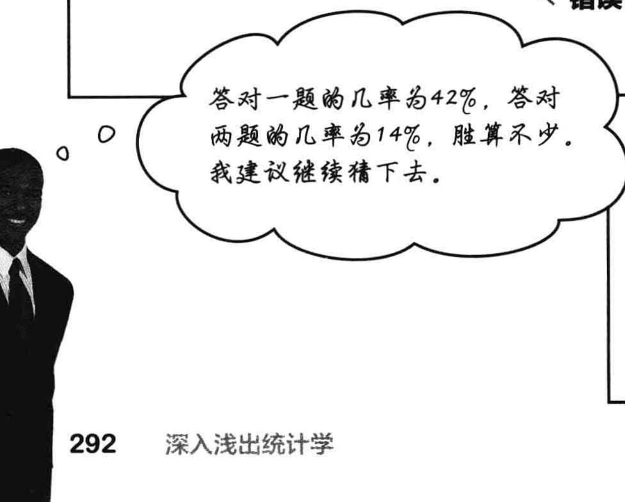
动动笔解答

这类问题的概率是多大？可以看出哪种模式？用X代表答对的题数，共3题。

x	P(X = x)	0.75的幂	0.25的幂
0	$0.75^3 = .422$	3	0
1	$3 \times 0.75^2 \times 0.25 = .422$	2	1
2	$3 \times 0.75 \times 0.25^2 = .141$	1	2
3	$0.25^3 = .015$	0	3

答对一题的情况有3种，每一种情况的概率均为0.752 × 0.25。





动动脑

请复习第6章“排列与组合”。你觉得对解决这类问题有帮助吗？

推广到求3个问题的概率

前面讲到了X的概率分布，X为答对的题数，总共3题。

与几何分布一样，这里的概率似乎也有某种模式。每一种概率都含有0.75和0.25的幂，随着x增大，0.75的幂减小，而0.25的幂增大。

一般，P(X = r)如下计算：

$$P(X = r) = \textbf{?} \times 0.25^{r} \times 0.75^{3-r}$$

r是答对的题数。

这是什么？

每道题的答对概率。

共3题。

每道题的答错概率。

即，为了求出答对r题的概率，可算出0.25^r，乘以0.75^{3-r}，然后将以上结果乘以某个数值。这个数值是多少呢？

缺少的数字是哪一个？

对于每一种概率，我们需要答对一定数目的问题，而答对一定数目的问题的方式不止一种。例如，总共3题，答对其中任意一题的情况有3种。还可以这样理解：存在3种不同的**组合**。

提醒一下：组合^nCr即从n个对象中选取r个对象的选取方式数目(不需要知道确切的选取顺序)。这正是我们现在碰到的情况，我们需要从3个问题中选取r个答对的问题。

第6章介绍了这种情况，必要时请复习一下。

即，3题中答对r题的概率可以这样计算：

$$P(X = r) = {}^3C_r \times 0.25^{r} \times 0.75^{3-r}$$

因此，根据这个公式，答对1题的概率为：

$$\begin{aligned} P(X = r) &= {}^3C_1 \times 0.25 \times 0.75^{3-1} \\ &= 3!/(3-1)! \times 0.25 \times 0.5625 \\ &= 6/2 \times 0.0625 \times 0.75 \\ &=0.422 \end{aligned}$$

这和上一页用图表算出来的结果一样。



动动笔解答

下面是第一轮的提问，都是关于游戏主持人的问题。

1. 他喜欢什么颜色？

- [ ] **A:** 红色
- [ ] **B:** 蓝色
- [ ] **C:** 绿色
- [x] **D:** 黄色

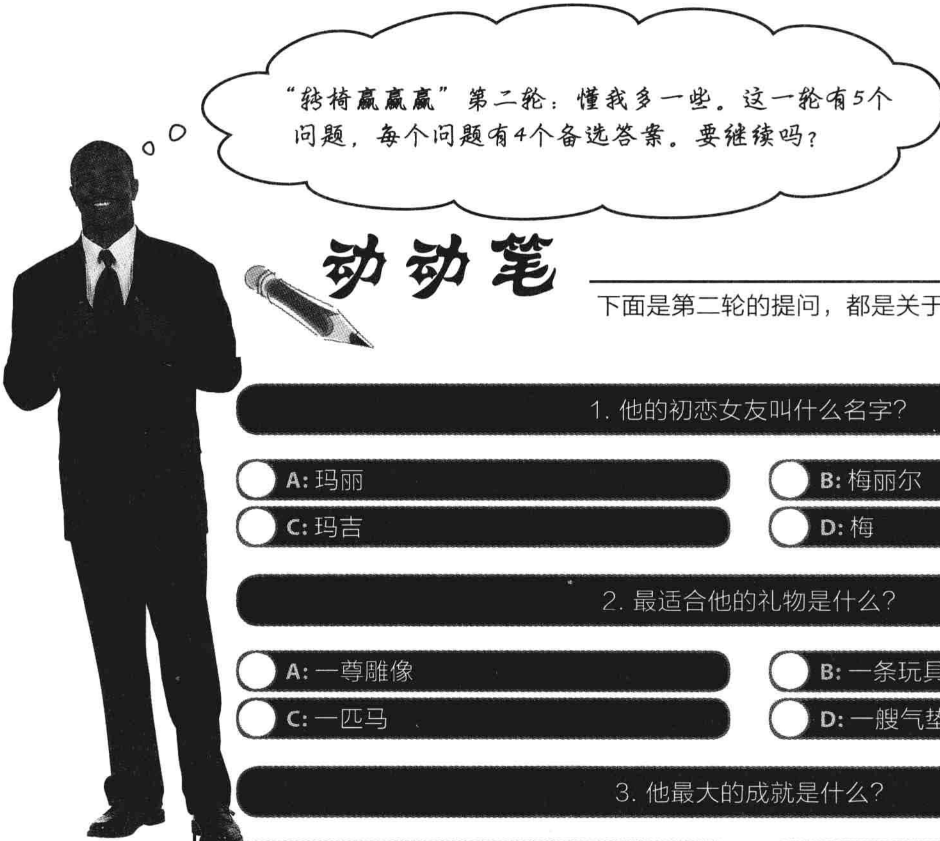
2. 他的生日在几月份？

- [ ] **A:** 1月
- [ ] **B:** 2月
- [x] **C:** 3月
- [ ] **D:**4月

3. 人们最喜欢他哪一点？

- [ ] **A:** 长相好看
- [ ] **B:** 有魅力
- [ ] **C:** 有幽默感
- [x] **D:** 机智





动动笔

下面是第二轮的提问，都是关于游戏主持人的问题。

1. 他的初恋女友叫什么名字？

- **A:** 玛丽
- **B:** 梅丽尔
- **C:** 玛吉
- **D:** 梅

2. 最适合他的礼物是什么？

- **A:** 一尊雕像
- **B:** 一条玩具狗
- **C:** 一匹马
- **D:** 一艘气垫船

3. 他最大的成就是什么？

- **A:** 主持智力节目
- **B:** 当选“2008年度统计邦先生”
- **C:** 为海豹保护区募得1000美元善款
- **D:** 发行唱片集

4. 他有什么不可告人的野心？

- **A:** 推出一系列体育设备
- **B:** 发行健身 DVD
- **C:** 推出自己的男装系列
- **D:** 推出自己的美发系列

5. 他在哪一年被外星人绑架了？

- **A:** 2005
- **B:** 2006
- **C:** 2007
- **D:** 2008

看来这些问题还是和上一轮一样难猜，所以还是要凭运气。

让我们看看能不能算出这些新问题的概率分布。

进一步推导概率算式

前面讲过，答对3个问题中的r个问题的概率是：

$$P(X = r) = {}^3C_r \times 0.25^r \times 0.75^{3-r}$$

其中，0.25为每道题的答对概率，0.75为每道题的答错概率。

第二轮“转椅赢赢赢”有5个问题，而不是3个。我们就不重新计算5个问题的解法了——让我们求出n个问题的解法，这样就能用同一个公式解决每一轮“转椅赢赢赢”的问题。

用n代替3就是了。

那么用哪个公式计算答对n个问题中的r个问题的概率呢？请看：

$$P(X = r) = {}^nC_r \times 0.25^r \times 0.75^{n-r}$$



没错，可以进行归纳。

设想每道题的答对概率是p，而每道题的答错概率是1-p，也就是q。答对n个问题中的r个问题的概率为：

$$P(X = r) = {}^nC_r \times p^r \times q^{n-r}$$

这类问题称为二**项分布**，让我们仔细看看。

二项分布细细看

猜测“转椅赢赢赢”各种问题的答案是**二项分布**的一个实例，二项分布包括下列条件：

1. 你正在进行一系列独立试验。
2. 每一次试验都存在失败和成功的可能，每一次试验的成功概率相同。
3. 试验次数有限。

这两个条件和几何分布的条件相同。

这个条件有变化。

和几何分布的情况一样，你要进行一系列独立试验，每一次试验结果或成功或失败。差别在于这一次你感兴趣的是获得成功的次数。

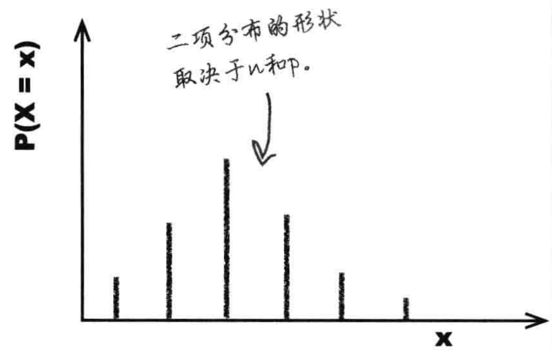
让我们用X表示“n次试验中的成功次数”，为了求出取得r次成功的概率，可用下列算式：

$$P(X = r) = {}^{n}C_{r}\, p^{r}\, q^{n-r} \qquad \text{其中} \qquad {}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

p是每一次试验的成功概率，n是试验次数。写作：

$$X \sim B(n, p)$$

根据n与p 的不同数值，二项分布的形状会发生变化，p越接近0.5，图形越对称。一般情况下，当p小于0.5时，图形向右偏斜；当p大于0.5时，图形向左偏斜。



期望和方差如何计算？

前面讲过如何使用二项分布计算基本概率，由此我们可以算出答对一定数目的问题的概率。但是，如果答案是随机选择的，那么我们到底能期望自己答对几个问题呢？算出期望可以帮助你作出更正确的选择，以便决定是否参加下一轮问题的回答。

让我们看看能否求出期望和方差的常规表达式。我们先算单次试验的期望和方差，然后看看是否能推广至n次独立的试验。

先看单次试验

假定我们只试验一次。每一次试验或是成功，或是失败，因此，在单次试验时，有可能取得0次或1次成功，如果$X \sim B(1,p)$，则成功1次的概率为p，成功0次的概率为q。

这是X的概率分布，X符合$X \sim B(1,p)$。

x	0	1
P(X = x)	q	p

我们可以根据以上条件求出X的期望和方差，让我们先算期望。

$$E(X) = 0q + 1p$$
$$= p$$

$$Var(X) = E(X^2) - E(X)^2$$
$$= (0q + 1p) - p^2$$
$$= p - p^2$$
$$= p(1 - p)$$
$$= pq$$

$E(X) = p$, so $E(X)^2 = p^2$

$E(X^2)$

因此，单次试验的$E(X) = p$，$Var(X) = pq$。那么n次试验呢？

动动脑

一般情况下，如果有n个独立观察结果，那么期望和方差是多少？在本例中对我们有何帮助？

奇妙池



让我们看看你是否能推导出$Y \sim B(n, p)$的期望和方差。你的任务是从奇妙池中捞出公式因子，将这些因子放入计算式中的横线上。每个因子只能用一次，不必使用所有因子。

提示：每个X_i是一次单独的试验。$E(X_i)= p$，$Var(X_i) = pq$。

你需要求出n个独立试验的期望和方差。

$E(X) = E(X_1) + E(X_2) + \cdots + E(X_n)$

$= \ldots\ldots\ldots\ E(X_i)$

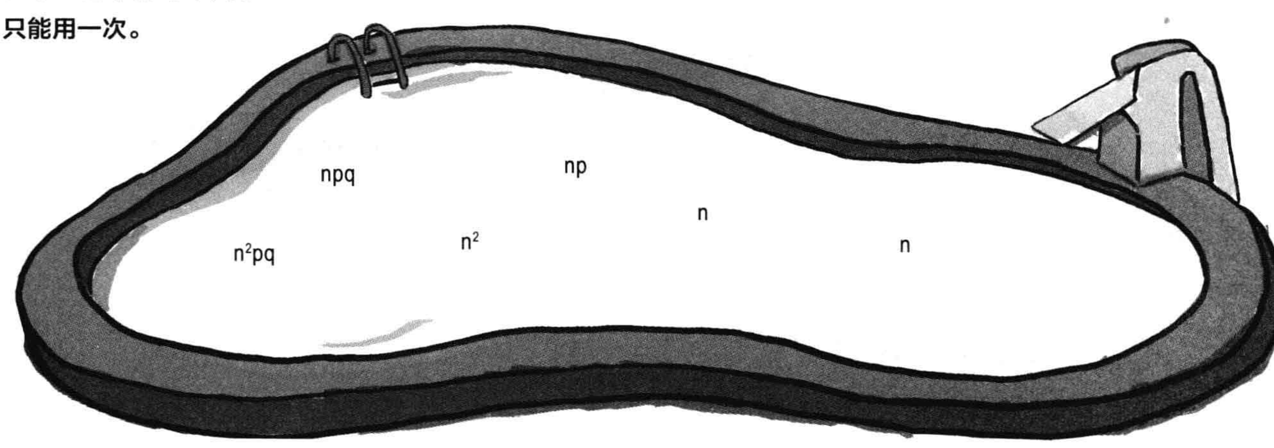
$= \ldots\ldots\ldots\ldots$

$Var(X) = Var(X_1) + Var(X_2) + \cdots + Var(X_n)$

$= \ldots\ldots\ldots Var(X_i)$

$= \ldots\ldots\ldots\ldots$

注意：池中的每个因子只能用一次。



奇妙池解答



让我们看看你是否能推导出Y ~ B(n, p)的期望和方差。你的任务是从奇妙池中捞出公式因子，将这些因子放入计算式中的横线上。每个因子只能用一次，不必使用所有因子。

提示：每个X_i是一次单独的试验。$E(X_i) = p$，$Var(X_i) = pq$。

你需要求出n个独立试验的期望和方差。

$$E(X) = E(X_1) + E(X_2) + ... + E(X_n)$$

$$= n\ E(X_i)$$

$$= np$$

由于试验是独立的，因此，$E(X_1) = E(X_2) = E(X_3)$，以此类推。

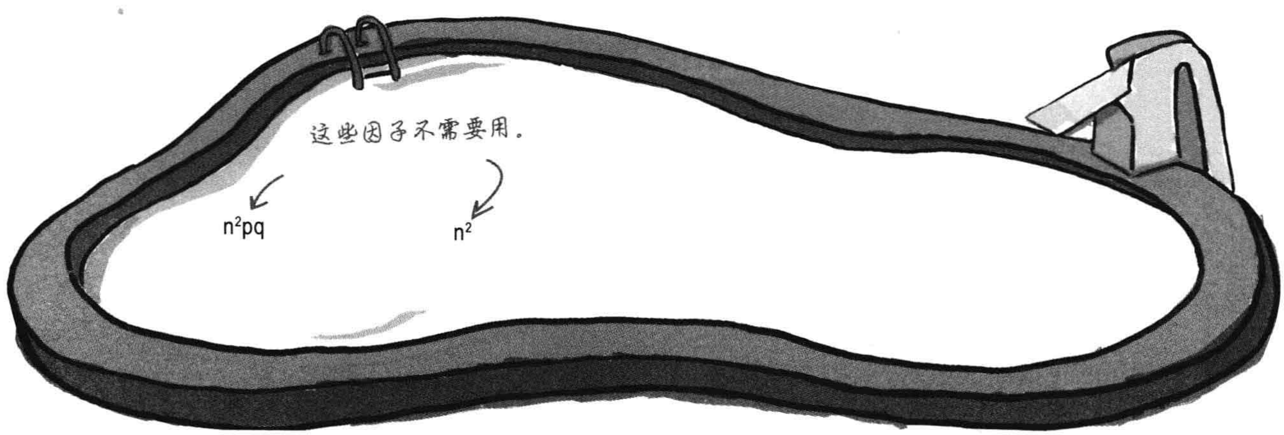
如果$X \sim B(n, p)$，则$E(X) = np$，$Var(X) = npq$

$$Var(X) = Var(X_1) + Var(X_2) + ... + Var(X_n)$$

$$= n\ Var(X_i)$$

$$= npq$$

由于试验是独立的，因此$Var(X_1) = Var(X_2) = Var(X_3)$，以此类推。



二项分布的期望与方差

让我们归纳一下前面做过的分析。首先看单次试验的情况：单次试验的成功概率为p，符合二项分布。根据这些条件，我们求出了单次试验的期望和方差。

然后我们分析了n个独立试验的情况，并利用简便方法求出了n次试验的期望与方差。我们发现，只要X ~ B(n, p)，则：

$$E(X) = np$$

$$Var(X) = npq$$

这些公式对所有二项分布都成立。

得出这个结论十分有用，因为这样一来，我们不用大量计算单个概率，就能迅速求出任何二项概率分布的期望和方差。

世上没有傻问题

问： 几何分布和二项分布看着很相似。它们有区别吗？分别应该在什么时候用呢？

答： 几何分布和二项分布确实有共同之处，二者处理的都是独立试验，每次试验都或是成功，或是失败。差别在于实际上要求的结果。在哪种情况下使用哪种概率分布取决于要求的结果。

如果试验次数固定，求成功一定次数的概率，则需要使用二项分布；使用二项分布还可以求出在n次试验中能够期望取得的成功次数。

如果你感兴趣的是在取得第一次成功之前需要试验多少次，则需要使用几何分布。

问： 几何分布是有众数的，二项分布有众数吗？

答： 有的。一个概率分布的众数就是具有最高概率的数值，如果p为0.5且n为偶数，则众数为np；如果p为0.5且n为奇数，则该概率分布有两个众数，即位于np左右两侧的两个数值。对于其他n值和p值，则需要通过反复试算的方法求众数，但一般都非常接近np。

问： 几何分布和二项分布都要进行大量试验，每一次试验的成功概率都必须一样吗？

答： 为了能应用几何分布和二项分布，每一次试验的成功概率都必须相同。如果不满足这个条件，则无论是几何分布还是二项分布都不适用。

问： 我试着算出了E(X)，但所得结果不是概率分布中的数值。我哪里做错了吗？

答： 计算E(X)的时候，结果有可能不是概率分布中的可能数值，即，结果有可能不是一个会实际出现的数值。得出这样的结果并不表示你算错了，别担心。

问： 还有其他类型的概率分布吗？

答： 有。接着读吧，更多的内容在等着你。

二项分布简明指南

下面是有关二项分布的简明总结，你可能用得上：

何时使用二项分布？

进行次数固定的独立试验时可使用二项分布，这时，每一次试验都存在成功或失败的可能，而你感兴趣的是成功或失败的次数。

如何计算概率？

公式为：

$$P(X = r) = {}^{n}C_{r}\, p^{r} q^{n-r}$$

$$ {}^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

其中p为单次试验的成功概率，q = 1 - p，n为试验次数，X为在n次试验中取得的成功次数。

期望和方差如何计算？

$$E(X) = np \qquad Var(X) = npq$$



练习解答

最后一轮“转椅赢赢赢”游戏中共有5个问题，每一题的答对概率是0.25。

1. 答对两题的概率是多少?

2. 答对3题的概率是多少?

3. 答对两题或3题的概率是多少?

4. 一题也答不对的概率是多少?

5. 期望和方差是多少?



最后一轮“转椅赢赢赢”游戏中共有5个问题，每一题的答对概率是0.25。

1. 答对两题的概率是多少?

如果X代表答对的题数，则$X \sim B(n, p)$：

$$P(X = 2) = {}^5C_2 \times 0.25^2 \times 0.75^3$$

$$= \frac{5!}{3!2!} \times 0.0625 \times 0.421875$$

$$= 10 \times 0.0264$$

$$= 0.264$$

2. 答对3题的概率是多少?

$$P(X = 3) = {}^5C_3 \times 0.25^3 \times 0.75^2$$

$$= \frac{5!}{2!3!} \times 0.015625 \times 0.5625$$

$$= 10 \times 0.00879$$

$$= 0.0879$$

3. 答对两题或3题的概率是多少?

$$P(X = 2 \text{ 或 } X = 3) = P(X = 2) + P(X = 3)$$

$$= 0.264 + 0.0879$$

$$= 0.3519$$



4. 一题也答不对的概率是多少?

$$P(X = 0) = 0.75^5$$

$$= 0.237$$

5. 期望和方差是多少?

$$E(X) = np$$

$$= 5 \times 0.25$$

$$= 1.25$$

$$Var(X) = npq$$

$$= 5 \times 0.25 \times 0.75$$

$$= 0.9375$$

动动笔解答

下面是第二轮的提问，都是关于游戏主持人的问题。

1. 他的初恋女友叫什么名字？

- A: 玛丽
- B: 梅丽尔
- ✓ C: 玛吉
- D: 梅

2. 最适合他的礼物是什么？

- A: 一尊雕像
- ✓ B: 一条玩具狗
- C: 一匹马
- D: 一艘气垫船

3. 他最大的成就是什么？

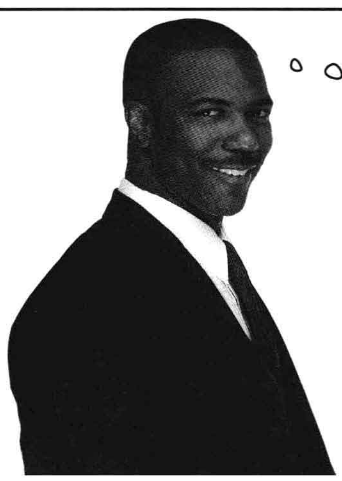
- A: 主持智力节目
- B: 当选“2008年度统计邦先生”
- ✓ C: 为海豹保护区募得1000美元善款
- D: 发行唱片集

4. 他有什么不可告人的野心？

- A: 推出一系列体育设备
- B: 发行健身DVD
- C: 推出自己的男装系列
- ✓ D: 推出自己的美发系列

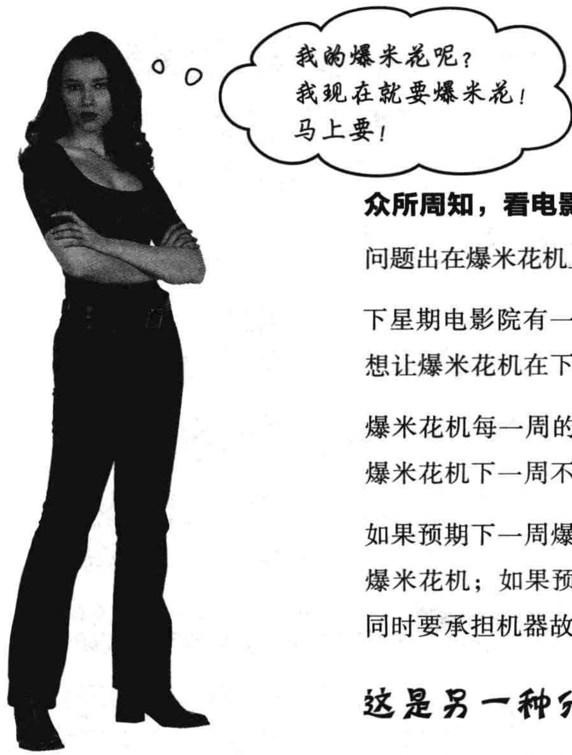
5. 他在哪一年被外星人绑架了？

- ✓ A: 2005
- B: 2006
- C: 2007
- D: 2008



您能参加这次智力游戏真是太好了，希望您下次再来。对了，统计邦电影院刚才给我们来了一个电话，爆米花出了点问题？

统计邦电影院遇到了问题



众所周知，看电影怎么少得了爆米花

问题出在爆米花机上，统计邦电影院的爆米花机总是坏，顾客们很不高兴。

下星期电影院有一个大型促销，影院经理希望一切都完美无缺。他可不想让爆米花机在下星期坏掉，否则就再也没人来看电影了。

爆米花机每一周的平均故障次数为3.4，或者说爆米花机的故障率为3.4。爆米花机下一周不发生故障的概率有多大？

如果预期下一周爆米花机会发生多次故障，则统计邦电影院会买一台新爆米花机；如果预期不会发生故障，他们将继续使用现在这台机器，但同时要承担机器故障的风险。

这是另一种分布

这次的问题与我们前面遇到过的问题不同。

这一次不存在一系列的试验，相反，这一次的情况是这样的：已知故障的发生几率，且该故障是随机发生的。

那么我们如何求出概率？

这一类问题的难点在于，尽管我们知道爆米花机每周的平均故障次数，但实际的故障次数却不是固定的。从总体上看，我们可以期望的故障次数是每周3或4次，但在倒霉的某一周，故障会多得多，而在顺利的某一周，故障则根本不会发生。

我们需要求出爆米花机下周不发生故障的概率。

听起来挺难吧？别担心，有一种概率分布是专门用来应付这种情况的，叫做**泊松分布**。

泊松分布细细看

泊松分布包括以下条件：

1. 单独事件在给定区间内随机、独立地发生，给定区间可以是时间或空间，例如可以是一个星期，也可以是一英里。

2. 已知该区间内的事件平均发生次数（或者叫做发生率），且为有限数值。该事件平均发生次数通常用希腊字母 λ (lambda) 表示。

让我们用X表示给定区间内的事件发生次数，例如一个星期内的损坏次数。如果X符合泊松分布，且每个区间内平均发生 λ 次，或者说发生率为 λ，则写作：

$$X \sim Po(\lambda)$$

我们就不在这里进行推导了。在求给定区间内发生r次事件的概率时，请使用下式进行计算：

$$P(X = r) = \frac{e^{-\lambda} \lambda^r}{r!}$$

别被表面现象吓住了，实际上计算方法十分简单直接。

e是一个数学常数，一般为2.718，只要把这个数字代入泊松分布公式就行了。许多科学计算器都有ex键，可以用这个键计算e的幂。

这个求概率的公式用到了指数函数e^x，x是未知数。大部分计算器都有这个标准函数，因此虽然这个公式看起来很复杂，实际应用却非常简单。

例如，如果$X \sim Po(2)$，则：

$$P(X = 3) = \frac{e^{-2} \times 2^3}{3!}$$

使用这个公式，代入r = 3，λ = 2。

$$= \frac{e^{-2} \times 8}{6}$$

$$= e^{-2} \times 1.333$$

$$= 0.180$$

那么，如果X符合泊松分布，其期望和方差如何？答案比你想像的可能要简单一些……

泊松分布的期望和方差

求泊松分布的期望和方差比求其他分布的期望和方差更容易。

如果X ~ Po(λ)，则E(X)为我们在给定区间内能够期望的事件发生次数，对于爆米花机来说，则为我们在普通的一周内能够期望的机器损坏次数，也就是说，E(X)是给定区间内的事件平均发生次数。

现在，如果X~Po(λ)，则事件平均发生次数以 λ 表示，即E(X)等于 λ ，这个参数决定了我们的泊松分布。



泊松分布（相较其他分布）更简洁的地方在于，它的方差也是 λ ，因此，如果X ~ Po(λ)，则：

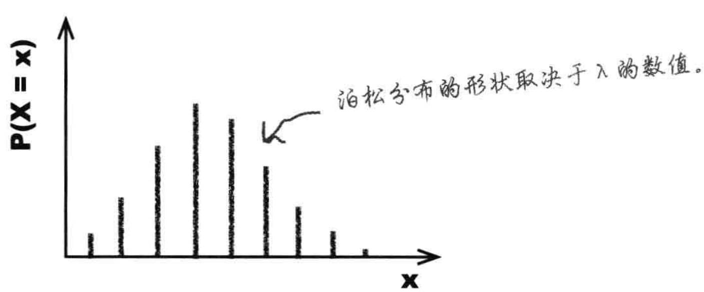
$$E(X) = \lambda \qquad Var(X) = \lambda$$

即，如果给你一个泊松分布Po(λ)，你根本不用做任何计算就能得出期望和方差 —— 泊松分布的参数本身就是期望和方差。

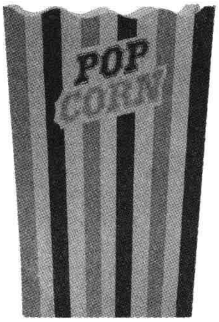
泊松分布是何形状？

泊松分布的形状随着 λ 的数值发生变化。λ 小，则分布向右偏斜，随着 λ 变大，分布逐渐变得对称。

如果 λ 是一个整数，则有两个众数， λ 和 λ -1，如果 λ 不是整数，则众数为 λ 。



化身爆米花机



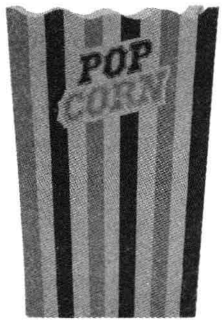
你的任务是假装自己是爆米花机，并说出你在下一周的一个特定时间段内发生故障的概率。记住，你发生损坏的平均次数是每周3.4次。

1. 下一周爆米花机不发生故障的概率是多少？

2. 下一周爆米花机发生3次故障的概率是多少？

3. 爆米花机发生故障的期望和方差是多少？

化身爆米花机解答



你的任务是假装自己是爆米花机，并说出你在下一周的一个特定时间段内发生故障的概率。记住，你发生损坏的平均次数是每周3.4次。

让我们用X代表爆米花机在一周内的故障次数，已知

$X \sim Po(3.4)$

1. 下一周爆米花机不发生故障的概率是多少？

如果不发生故障，则X必须为0。

$$P(X = 0) = \frac{e^{-\lambda}\lambda^{r}}{r!}$$

$$= \frac{e^{-3.4} \times 3.4^{0}}{0!}$$

$$= \frac{e^{-3.4} \times 1}{1}$$

$$= 0.033$$

看来我们可以期望爆米花机在下周只发生3.4次故障，所以我们可以冒险不买新爆米花机——可别告诉那些看电影的。



2. 下一周爆米花机发生3次故障的概率是多少？

$$P(X = 3) = \frac{e^{-3.4} \times 3.4^{3}}{3!}$$

$$= \frac{e^{-3.4} \times 39.304}{6}$$

$$= 0.033 \times 6.55$$

$$= 0.216$$

3. 爆米花机发生故障的期望和方差是多少？

$E(X) = \lambda$

$= 3.4$

$Var(X) = \lambda$

$= 3.4$

世上没有傻问题

问： 为什么用λ代表泊松分布的均值？为什么不像以前一样用μ呢？

答： 这是因为泊松分布的分布参数、期望和方差全都相等，因此用λ，这样可以确保公正。

问： 泊松分布的公式是怎么来的？

答： 实际上可以从其他公式推导出来，但会涉及很多数学知识。在实际应用中，最好的做法是记住这个公式及其应用条件。

问： 泊松分布和其他概率分布有何差别？

答： 主要差别是泊松分布不需要做一系列试验，但它描述了事件在特定区间内的发生次数。

问： λ必须是整数吗？

答： 完全不是这样。λ可以是任何非负数，但不能是负数，因为它代表一定区间内的事件平均发生次数，事件发生次数为负数是没有意义的。

问： 公式中的“e”到底是什么意思？

答： e是一个数学常数，即数字2.718，在计算泊松分布时，要在公式中代入常数2.718。

常数e在微积分中应用频繁，广泛用于从计算复利到高等概率理论的各种应用。对e的深入讨论不在本书范围内。

问： 我用泊松分布计算概率的时候经常出错，哪里容易引发错误？

答： 有两个部分容易搞错。第一，一定要用对公式，r和λ很容易混淆，因此一定要确保二者正确无误。

第二，一定要在算式中正确应用e^x函数，把$e^{-\lambda}$留到最后再算是一个办法——即先把其他东西算出来，最后再乘以$e^{-\lambda}$。



统计邦电影院又碰到了一个问题。

不仅爆米花机总是出故障，现在，连饮料机也开始出故障了。饮料机每周的平均故障次数是2.3。

下个星期就要大促销了，任何机器坏了影院经理都要吃不了兜着走。下个星期，爆米花机和饮料机都不出故障的概率有多大？

动动脑

饮料机的概率分布是怎样的？我们如何求出爆米花机和饮料机在下个星期都不出故障的概率？

概率分布是怎样的？

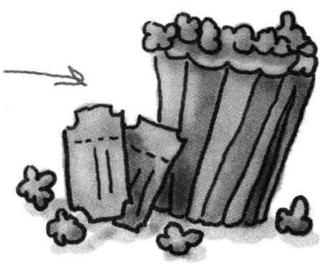
让我们好好看看这种情况。

我们有两种机器：爆米花机和饮料机，每种机器在一周内的平均故障次数已经知道，求下一周机器不出故障的概率。

下面是两种机器的分布：

爆米花机

爆米花机每周发生故障的平均次数是3.4。



X ~ Po(3.4)

饮料机

饮料机每周发生故障的平均次数是2.3。



Y ~ Po(2.3)

如果X代表爆米花机每周发生故障的次数，Y代表饮料机每周发生故障的次数，则X和Y都符合泊松分布，另外，X和Y是相互独立的，即爆米花机是否发生故障对饮料机发生故障的概率没有影响，而饮料机是否发生故障也对爆米花机发生故障的概率没有影响。

我们需要求出下个星期故障总次数为0的概率，即：

$$P(X + Y = 0)$$

动动脑

回头复习概率章节，如果X和Y是独立变量，那么如何求X+Y的概率？

组合泊松变量

前面的章节中讲过，如果X和Y是独立随机变量，则：

$$P(X + Y) = P(X) + P(Y)$$

$$E(X + Y) = E(X) + E(Y)$$

即如果$X \sim Po(\lambda_x)$ 且 $Y \sim Po(\lambda_y)$，则：

$$\mathbf{X + Y \sim Po(\lambda_x + \lambda_y)}$$

即，如果X和Y都符合泊松分布，则X+Y也符合泊松分布。也就是说，可以利用X和Y的分布情况求出X+Y的概率。



动动笔

如果X是爆米花机的故障次数，Y是饮料机的故障次数，则 $X \sim Po(3.4)$，$Y \sim Po(2.3)$。

1. X+Y的分布情况如何？

2. 求出X+Y的分布后，可以根据分布求出概率。$P(X + Y = 0)$是多少？

动动笔解答

如果X是爆米花机的故障次数，Y是饮料机的故障次数，则X~ Po(3.4)，Y ~ Po(2.3)。

1. X+Y的分布情况如何？

$$\lambda_x + \lambda_y = 3.4 + 2.3$$
$$= 5.7$$
$$X + Y \sim Po(5.7)$$

2. 求出X+Y的分布后，可以根据分布求出概率。P(X + Y = 0)是多少？

$$P(X + Y = 0) = \frac{e^{-\lambda}\lambda^r}{r!}$$
$$= \frac{e^{-5.7} \times 5.7^0}{0!}$$
$$= \frac{e^{-5.7} \times 1}{1}$$
$$= 0.003$$



世上没有傻问题

问： 这是不是说前面学过的关于概率和期望的简明算法也适用于泊松分布？

答： 不错。由于爆米花机是否发生故障对饮料机发生故障的概率没有影响，反过来，饮料机是否发生故障对爆米花机发生故障的概率也没有影响，因此，X和Y都是独立随机变量，于是所有适用于独立变量的简明计算方法都能为我们所用。

问： X+Y为什么会符合泊松分布？

答： 这是因为X和Y都是独立变量，且都符合泊松分布。

爆米花机和饮料机都会随机出现故障，但有一个平均故障率，这意味着将两种机器放在一起后，也会随机发生故障，也会有一个平均故障率，也就是两种机器合起来仍然符合泊松分布的条件。

问： 所以我们就能像应用其他泊松分布一样应用X+Y的分布？

答： 是的，我们可以用完全相同的方式对待X+Y的分布，因此，只要知道参数λ，就能求出概率。

案件：破碎的饼干

凯特在统计邦曲奇饼厂工作，她的工作是确保每一盒饼干都符合工厂严格的质量要求。凯特知道每块饼干发生破碎的概率为0.1，她的老板要她求出一盒容量为100块饼干的盒子里出现15块碎饼干的概率。“这容易”，她说道，“用二项分布计算好了，n为100，p为0.1。”



凯特拿出计算器，可当她计算100！的时候，计算机显示出错，因为数字太大。“哦，”老板说，“你只好用手工方法计算了。我现在可是要回家了，祝你晚上愉快。”

凯特瞪着计算器，动起了脑筋。随后她笑了，“也许我今晚可以早点走，到底还是有办法的。”

不出1分钟，凯特就算出了要求的概率。她设法绕过了100！的计算，求出了概率。她拿起外套走出了厂门。

凯特怎么能这么快就避开计算器的限制算出概率？

伪装下的泊松分布

泊松分布还有一个用途：在特定条件下可以用来近似代替二项分布。



有时候，使用泊松分布比使用二项分布更简单

例如，假设需要计算一个二项概率，其中n为3000。在此过程中需要计算3000！，就算有一个好计算器，这也很难计算出来。因此，懂得用泊松分布正确地求解近似答案就显得十分有用。

那么我们在什么条件下能用这种近似法，该如何用？

假设我们有一个变量X，且X~B(n,p)，要求有这样一种条件：B(n, p)近似等于Po(λ)。

让我们首先研究两种分布的期望和方差。我们的目标是找出泊松分布的期望和方差近似等于二项分布的期望和方差的情况，即希望：

期望 → **λ** 近似 **np**

方差 → **λ** 近似 **npq**

→ **np** 近似 **npq**

当q近似等于1且n很大时，np和npq近似相等。即：

当n很大且p很小时，可以用X ~ Po(np)近似代替X ~ B(n, p)。

当n大于50且p小于0.1时，为典型的近似情况。

练习

一个学生要参加一场考试，但他没有做任何复习。他需要猜测每一题的答案，每一题的答对概率是0.05。考卷上共有50个问题，他答对5题的概率是多少？用二项分布的泊松分布近似法求解。

世上没有傻问题

问： 为什么有时候需要用泊松分布近似代替二项分布进行计算？

答： 当n很大时，计算$^{n}C_{r}$比较困难，有些计算器会发生内存不足的情况，且太大的计算结果会难以处理。使用泊松分布进行近似计算可以克服以上困难。

问： 那么什么时候可以使用这种近似法？

答： 当n很大(比如大于50)，p很小(比如小于0.1)，这时可以使用近似法，在这种情况下，二项分布和泊松分布近似相等。

问： 为什么把np作为泊松分布的参数？

答： 泊松分布只有一个参数λ，且E(X)= λ。这就是说，如果我们将泊松分布作为二项分布的近似，则可以代入二项分布的期望np。



一个学生要参加一场考试，但他没有做任何复习。他需要猜测每一题的答案，每一题的答对概率是0.05。考卷上共有50个问题，他答对5题的概率是多少？用二项分布的泊松分布近似法求解。

让我们用X表示学生猜对的问题的数目，在本例中，$n=50$，$p=0.05$，$np=2.5$，于是可以用$X \sim Po(2.5)$近似计算概率。

$$P(X=5)=\frac{e^{-\lambda}\lambda^{r}}{r!}$$

$$=\frac{e^{-2.5}\times 2.5^{5}}{5!}$$

$$=\frac{e^{-2.5}\times 97.65625}{120}$$

$$=e^{-2.5}\times 0.8138$$

$$=0.067$$

破案：破碎的饼干

凯特怎么能这么快就避开计算器溢出错误算出概率？



凯特发现，尽管需要用二项分布进行计算，但n和p的数值却允许她用泊松分布对概率进行近似计算。许多计算器无法计算大阶乘，有时候这会令二项分布无法作为，这时懂得用泊松分布进行近似计算会大大节省你的时间。

有人要爆米花吗？



本章内容已经接近尾声，通过学习三种最重要的离散概率分布，你的概率和统计知识又长进了不少。你深入了解了概率分布的作用，掌握了既能节省时间、又能得出可靠结果的简明算法，这些技术将在本书后续章节发挥作用。

小坐一会儿，吃点儿爆米花吧，犒劳犒劳自己。

泊松分布简明指南

下面是有关泊松分布的简明总结，你可能用得上：

何时使用泊松分布？

在遇到独立事件时(例如机器在给定区间内发生故障)，若已知 λ (即给定时间区间内的事件平均发生次数)且你感兴趣的是一个特定时间区间内的发生次数，这时可使用泊松分布。

如何计算概率、期望和方差？

计算方法如下：

$$P(X = r) = \frac{e^{-\lambda}\lambda^r}{r!} \qquad E(X) = \lambda \qquad Var(X) = \lambda$$

如何对独立随机变量进行组合？

如果$X \sim Po(\lambda_x)$且 $Y \sim Po(\lambda_y)$，则：

$$X + Y \sim Po(\lambda_x + \lambda_y)$$

泊松分布与二项分布有何关系？

如果$X \sim B(n, p)$，当n较大而p较小时，X可以近似表示为：

$$X \sim Po(np)$$



加强练习

下面是一些实例。你的任务是说出每个实例符合哪种概率分布，指出期望和方差，并求出各种概率。

1. 某人正在打保龄球，他击倒所有球柱的概率为0.3，如果他可以掷球10次，在3次以内击倒所有球柱的概率是多大?

2. 一辆公共汽车平均每15分钟会停一站。在15分钟以内不出现公共汽车的概率有多大?

3. 有20%的麦片盒里装有免费玩具，每盒一个。打开不到4只麦片盒就能得到第一个免费玩具的概率有多大?

加强练习解答

下面是一些实例。你的任务是说出每个实例符合哪种概率分布，指出期望和方差，并求出各种概率。

1. 某人正在打保龄球，他击倒所有球柱的概率为0.3，如果他可以掷球10次，在3次以内击倒所有球柱的概率是多大？

如果用X代表这个人击倒全部球柱的次数，则 $X \sim B(10, 0.3)$：

$$E(X) = np = 10 \times 0.3 = 3$$

$$Var(X) = npq = 10 \times 0.3 \times 0.7 = 2.1$$

通用概率 $P(X = r) = {}^nC_r \times p^r \times q^{n-r}$

$$P(X = 0) = {}^{10}C_0 \times 0.3^0 \times 0.7^{10} = 1 \times 1 \times 0.028 = 0.028$$

$$P(X = 1) = {}^{10}C_1 \times 0.3^1 \times 0.7^9 = 10 \times 0.3 \times 0.04035 = 0.121$$

$$P(X = 2) = {}^{10}C_2 \times 0.3^2 \times 0.7^8 = 45 \times 0.09 \times 0.0576 = 0.233$$

$$P(X < 3) = P(X = 0) + P(X = 1) + P(X = 2) = 0.028 + 0.121 + 0.233 = 0.382$$

2. 一辆公共汽车平均每15分钟会停一站。在15分钟以内不出现公共汽车的概率有多大？

如果用X表示每15分钟以内停靠的公共汽车的数量，则 $X \sim Po(1)$。

$E(X) = \lambda$
$= 1$

$Var(X) = \lambda$
$= 1$

通用概率 $P(X = r) = \dfrac{e^{-\lambda}\lambda^r}{r!}$

$P(X = 0) = \dfrac{e^{-1} \times 1^0}{0!}$

$= \dfrac{e^{-1} \times 1}{1}$

$= 0.368$

3. 有20%的麦片盒里装有免费玩具，每盒一个。打开不到4只麦片盒就能得到第一个免费玩具的概率有多大？

如果用X表示为了找出第一个玩具需要打开的麦片盒的数目，则 $X \sim Geo(0.2)$。

$E(X) = 1/p$
$= 1/0.2$
$= 5$

$Var(X) = q/p^2$
$= 0.8/0.2^2$
$= 0.8/0.04$
$= 20$

通用概率 $P(X \leq r) = 1 - q^r$

$P(X \leq 3) = 1 - q^r$
$= 1 - 0.8^3$
$= 1 - 0.512$
$= 0.488$

要 点

- **几何分布**的应用条件：进行一系列独立试验，每一次试验或成功或失败，每一次试验的成功概率相同，你主要想知道的是：为了取得第一次成功，需要进行多少次试验。

- 如果符合几何分布的条件，那么用X表示为了取得第一次成功需要试验的次数，用p代表单次试验的成功概率，则：

  $X \sim Geo(p)$

- 如果$X \sim Geo(p)$，则下列概率算式成立：

  $P(X = r) = pq^{r-1}$

  $P(X > r) = q^r$

  $P(X \leqslant r) = 1 - q^r$

- 如果$X \sim Geo(p)$，则：

  $E(X) = 1/p$

  $Var(X) = q/p^2$

- **二项分布**的应用条件：进行一系列次数有限的独立试验，每一次试验或成功或失败，每一次试验的成功概率相同，你主要想知道的是：在n次试验中能成功多少次。

- 如果符合二项分布的条件，那么用X表示n次试验中的成功次数，用p代表单次试验的成功概率，则：

  $X \sim B(n, p)$

- 如果$X \sim B(n, p)$，则可通过下式计算概率：

  $P(X = r) = {}^nC_r\, p^r q^{n-r}$

其中：

$${}^nC_r = \frac{n!}{r!\,(n-r)!}$$

- 如果$X \sim B(n, p)$，则：

  $E(X) = np$

  $Var(X) = npq$

- **泊松分布**的应用条件：单个事件在给定区间内随机、独立地发生，已知给定区间内的事件平均发生次数，或者叫发生率，且这个发生次数或发生率是有限的，主要想知道的是：给定区间内的事件发生次数。

- 如果符合泊松分布的条件，那么用X表示给定区间内的事件发生次数，用λ代表发生率，则：

  $X \sim Po(\lambda)$

- 如果$X \sim Po(\lambda)$，则：

  $$P(X = r) = \frac{e^{-\lambda}\ \lambda^r}{r!}$$

  $E(X) = \lambda$

  $Var(X) = \lambda$

- 如果$X \sim Po(\lambda x)$，$Y \sim Po(\lambda y)$，且X和Y是独立的，则：

  $X + Y \sim Po(\lambda_x + \lambda_y)$

- 如果$X \sim B(n, p)$，其中n足够大，p足够小，则可将该分布近似看作$X \sim Po(np)$。

8 正态分布的运用

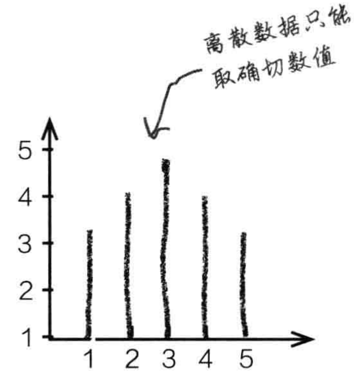
保持正态



离散概率分布并非无所不能。

到目前为止，我们接触到的都是可以指定确切数值的概率分布。然而并非所有数据集合都是如此，还有几类数据**并不符合**我们之前遇到的概率分布。我们将在这一章里讲解所谓的**连续型概率分布**，并介绍最重要的概率分布类型之一——**正态分布**。

离散数据可取确切值……



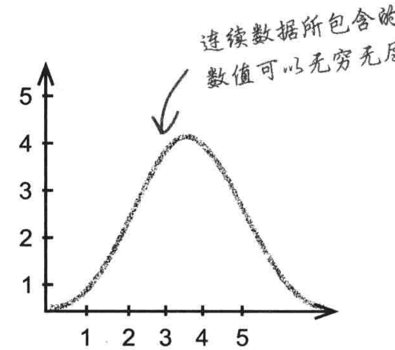
前面讲到的概率分布涉及的都是**离散数据**，即数据由一个个单独的数值组成，其中的每一个数值都有相应概率。例如，在分析老虎机收益概率分布时，每一局赌局可能出现的收益数额是确定的，我们很清楚各种情况的赔率，也知道自己有机会赢到其中一种。

如果是离散数据，则为数值型数据，只能取确切值。离散数据往往能以某种方式进行计数，例如糖果机中的糖果数目，智力游戏中答对的问题的数目，或是机器在一个特定时段内的故障次数。



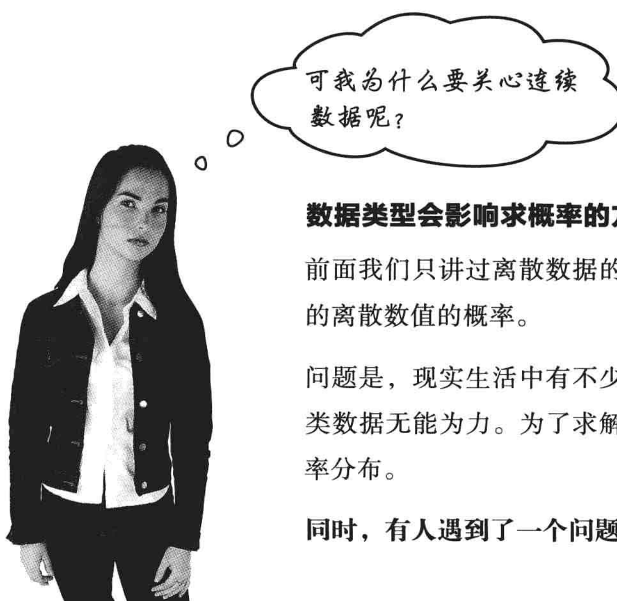
但并非所有数值型数据都是离散的

一一列举一个数据集中的所有数值并不总是能够实现。有时候，数据涵盖的是一个范围，这个范围内的任何一个数值都有可能成为事件结果。例如，假定有人让你精确地测量几段丝线的长度，并且已知这些丝线的长度在10英寸到11英寸之间，你的测量结果可能会是10英寸、10.1英寸、10.01英寸，等等，因为丝线长度可以是这个范围内的任意值。



这样的数据叫做**连续数据**，连续数据往往通过测量得到，而不是通过计数得到，测量结果在很大程度上取决于测量精度要求。





数据类型会影响求概率的方法。

前面我们只讲过离散数据的概率分布，利用这些概率分布，我们可以求出确切的离散数值的概率。

问题是，现实生活中有不少问题所牵涉到的都是连续数据，离散概率分布对这类数据无能为力。为了求解连续数据的概率，你需要懂得连续数据以及连续概率分布。

同时，有人遇到了一个问题……

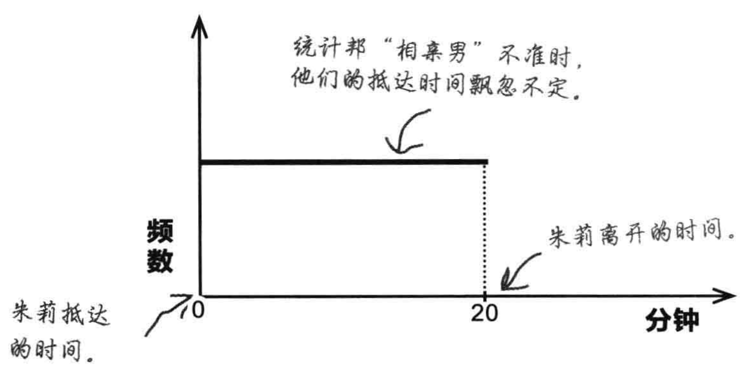
推迟几分钟？

朱莉是一名学生，她最好的朋友不停地安排她“相亲”，希望她能找到她的“他”。唯一的麻烦是，许多“相亲男”都不准时到场，或者根本就不现身。

朱莉讨厌孤零零地等待约会对象出现，于是她给自己立了规矩：如果等过20分钟对方还不来，她就离场。



下面这张频数图显示出朱莉为了见到约会者而等待的时间：



动动脑

我们需要求出朱莉为了见到约会对象而等待的时间的概率。这些时间量是离散的还是连续的？为什么？你认为我们该如何求出概率？

我们需要求连续数据的概率分布

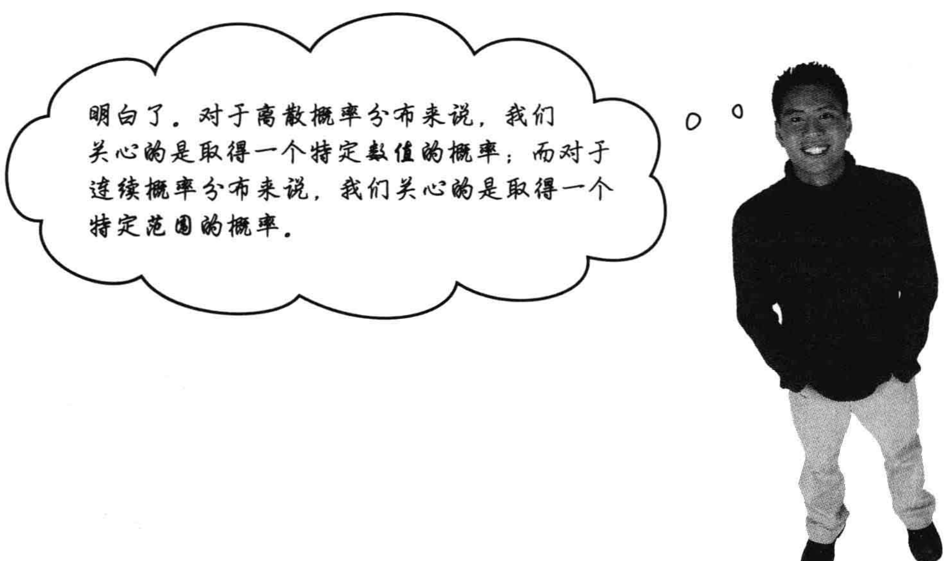
我们需要求出这种情况的概率：朱莉为了见到约会对象而等待5分钟以上。问题是，朱莉的等待时间是连续数据，也就是说，我们前面学过的概率分布在这里不适用。

处理离散数据时，我们可以找出特定的概率分布。为此，我们可以将每个数值的概率列于表格，也可以指出数据符合某个特定概率分布(例如二项分布或泊松分布)，通过这些做法，可以确定每一个可能数值的概率。例如，在我们求出肥蛋赌场每一台老虎机的每局收益概率分布后，我们就知道所有可能赢得的金额，还能算出每一种赢钱金额的概率。

对于离散数据，我们能给出每一个数值的概率。

x	-1	4	9	14	19
P(X = x)	0.977	0.008	0.008	0.006	0.001

连续数据则是另一番情形。我们再也无法给出每一个数值的概率，因为我们不可能列举每一个精确数值。例如，朱莉的约会者可能会在4分钟以后出现，在4分钟10秒以后出现，或在4分钟10.5秒以后出现，我们不可能数清楚所有的可能时间。相反，我们需要关心的是一个特定精度水平，以及取得一个数值**范围**的概率。

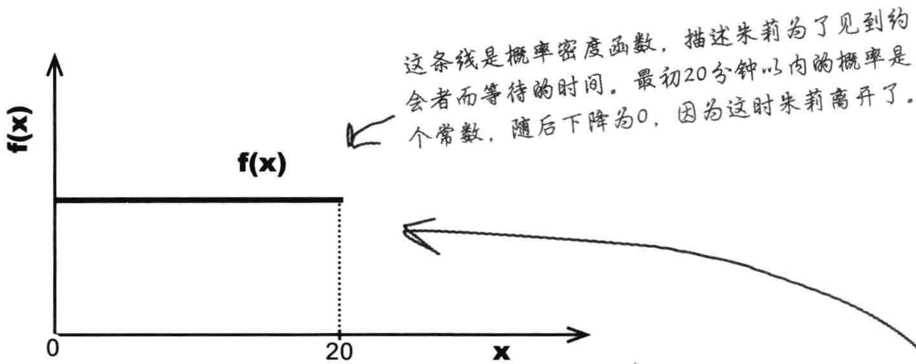


概率密度函数可用于描述连续数据

我们可以用**概率密度函数**描述连续随机变量的概率分布。

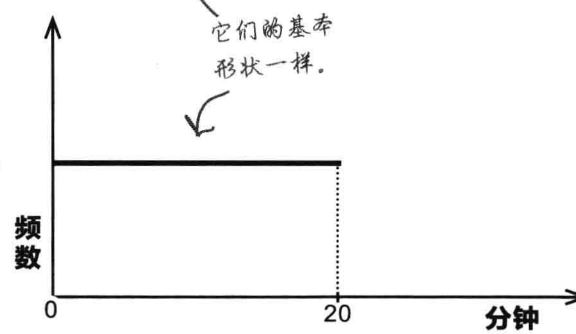
概率密度函数f(x)是这样一种函数：通过它可以求出一个数据范围内的某个连续变量的概率，它向我们指出该概率分布的形状。

下面是一张概率密度函数图，示意了朱莉为了见到约会者而等待的时间。



看出来了吗？这个图形与频数图形多么相符。这并非巧合。

概率的实质是告诉我们事情发生的可能程度，而频数告诉我们数值出现的频繁程度。频数越高，数值出现的概率越大。由于在最初20分钟内，朱莉的等待时间的频数为常数，这意味着概率密度函数也是常数。

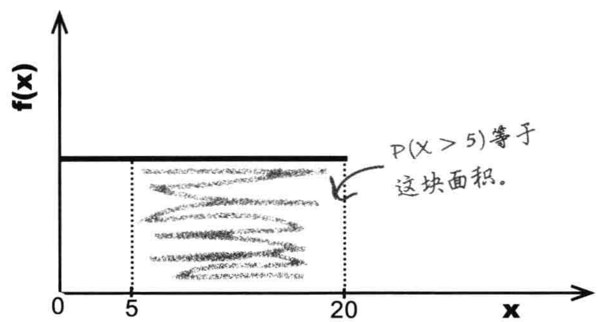




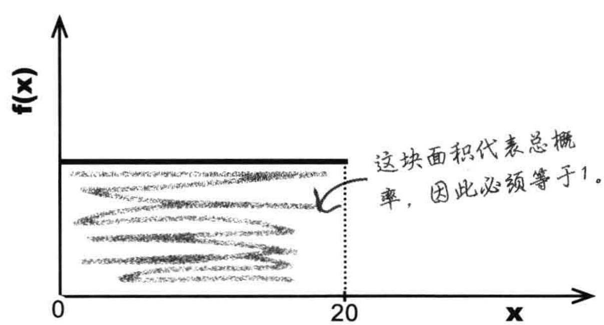
概率 = 面积

连续随机变量的概率通过面积表示。为了求出一个特定数值范围的概率，首先可画出概率密度函数，位于函数图形下方且介于这个特定数值范围之间的面积就是这个特定数值范围的概率。

例如，我们想求出朱莉为了见到约会对象而等待5−20分钟的概率，可以画出概率密度函数，再求出位于这个概率密度函数下方且x值介于5−20之间的面积。



线下总面积必须等于1，因为总面积代表总概率 —— 对于任何概率分布来说，总概率必须等于1，因此面积也必须等于1。



让我们利用这张图求出朱莉为了见到约会者而需要等待5分钟以上的概率。

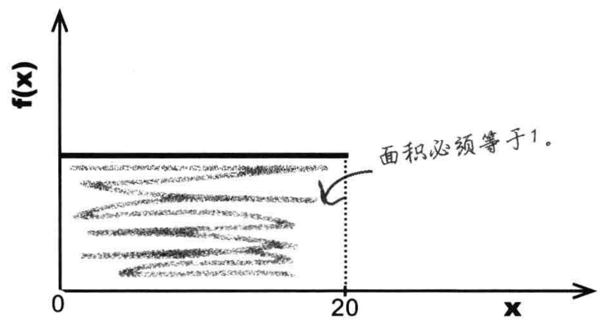
动动脑

线下总面积必须为1。f(x)的数值是多少？ 提示：是个常数。

欲算概率，先求f(x)……

在为朱莉算出概率之前，我们需要求得f(x)，即概率密度函数。

我们已经知道f(x)是一个常数，也知道这个函数下方的总面积等于1。观察f(x)的图形可知，图形下方是一个矩形，底宽为20。只要求出矩形的高，就可以得出f(x)的数值。



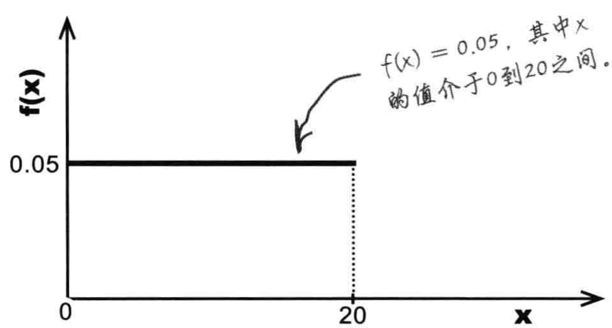
矩形的面积等于宽乘以高，即：

$$1 = 20 \times 高$$
$$高 = 1/20$$
$$= 0.05$$

这意味着f(x)必须等于0.05，才能确保线下面积等于1。即：

$f(x) = 0.05$　　其中x的值介于0到20之间。

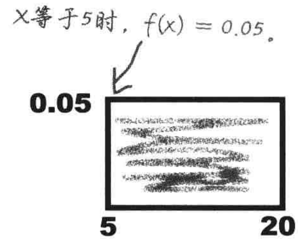
图形如下：



求出概率密度函数后，就可以求P(X > 5)了。

再求面积，可得概率

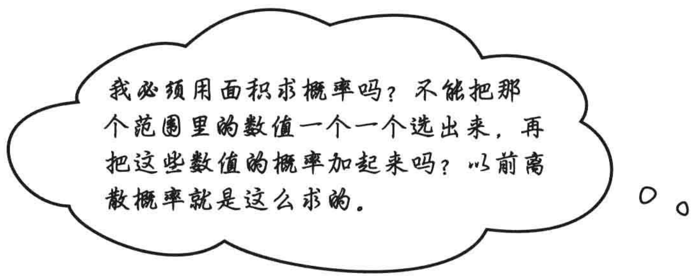
概率密度线下方介于5–20之间的区间是一个矩形，于是算出矩形面积将能得出概率P(X > 5)。



$$P(X > 5) = (20 - 5) \times 0.05$$
$$= 0.75$$

← 矩形面积 = 底 X 高

所以，朱莉等待5分钟以上的概率为0.75。





这种做法不适用于连续概率。

对于连续概率，我们**必须**通过计算概率密度曲线下方的面积得出概率。

不能通过把数值范围内的每一个数值的概率相加得出连续概率分布的概率，原因是数值个数无穷无尽，因此求和计算也会无休无止。

对于连续概率分布的概率，唯一的办法就是算出由连续概率函数形成的曲线下方的面积。

处理连续数据时，所计算的是一个数值范围的概率。

世上没有傻问题

问： 有一种函数叫做概率密度函数，那么什么是概率密度？

答： 概率密度指出各种范围内的概率的大小，通过概率密度函数进行描述。它与我们在第一章碰到过的频数密度十分相似。概率密度通过面积标示概率大小，而频数密度通过面积标示频数大小。

问： 难道概率密度和概率不是一回事？

答： 概率密度是一种表示概率的方法，但它并非概率本身。概率密度函数是图形中的一条线条，而概率则是这条线下方的一定数值范围内的面积。

问： 我明白了，这么说，如果有一张图，图中画出了概率密度函数，可以通过观察面积求出概率，而不是直接从图上读出概率。

答： 完全正确。对于连续数据，需要通过计算面积求出概率。从图上直接读出概率数值仅适用于离散概率的求解。

问： 必须通过计算面积求概率……这是不是搞复杂了？我是说，要是概率密度函数是一条曲线，而非直线，那该怎么办？

答： 还是行得通，但需要用到微积分，因此本书不打算让你进行这类计算。问题的关键是，要明白概率的来历，以及如何理解这种概率。

如果你**实在**对通过微积分计算概率感兴趣，无论如何都想试试，请大胆尝试，放手去学吧。

问： 关于概率范围，你已经讲过不少。我如何求出一个精确数值的概率？

答： 在处理连续数据的时候，实际上考虑的是一个可以接受的精度，并且基于这些数值形成一个范围。让我们看一个例子：

假定你想要一段丝线，长度10英寸，精确到英寸。虽然“你需要一段正好长10英寸的丝线”这种说法最容易脱口而出，但这并不完全正确。你**真正**想要的是一段长度介于9.5英寸到10.5英寸之间的丝线，因为你想让这段10英寸长的丝线“精确到英寸”。即，你想求出长度介于9.5英寸到10.5英寸这个范围内的概率。

问： 如果我想求某一个精确的数值的概率，会是多少？

答： 结果为0——猛一听可能会觉得有违直觉，但你的问题其实可以这样理解 ：求一个具有无穷小数位数的精确数值的概率。

让我们再以丝线长度为例：如果你需要一段长度正好等于10英寸的丝线，会出现什么局面？——你会需要用一台高倍放大镜，以原子大小为精度，量出一段10英寸长的丝线。

“丝线的长度正好为10英寸”这个事件基本上不可能发生。也就是其概率为零

问： 但我确信不需要那样高的精度。精确到百分之一英寸就够了，肯定是这样的，对吧？

答： 啊，这样就不是在讨论求一个具有无穷精度的数值的概率，而是回到10英寸长度的测量精度问题上了——你用自己选定的精度来构建可以接受的测量范围，得以算出概率。



化身概率密度函数

一些概率密度函数找不到它们的概率了，你的任务是假装自己是概率密度函数，算出指定数值范围内的概率。必要时可画图帮助。

1. $f(x) = 0.05$，其中$0 < x < 20$。
求$P(X < 5)$。

2. $f(x) = 1$，其中$0 < x < 1$。
求$P(X < 0.5)$。

3. $f(x) = 1$，其中$0 < x < 1$。
求$P(X > 2)$。

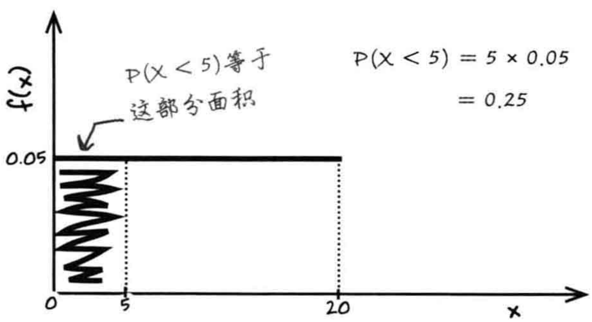
4. $f(x) = 0.1 - 0.005x$，其中$0 < x < 20$。
求$P(X > 5)$。



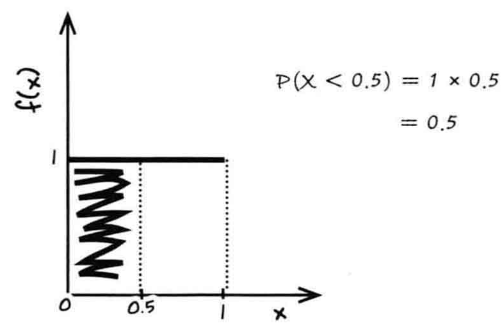
化身概率密度函数解答

一些概率密度函数找不到它们的概率了，你的任务是假装自己是概率密度函数，算出指定数值范围内的概率。必要时可画图帮助。

1. f(x) = 0.05，其中0 < x < 20。
求P(X < 5)。



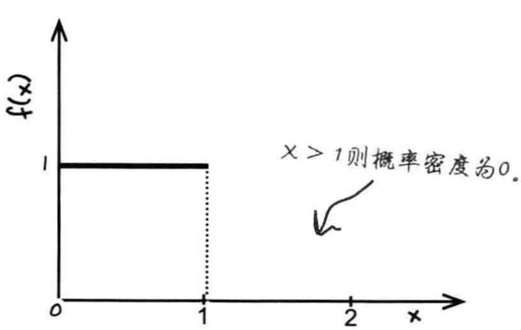
2. f(x) = 1，其中0 < x < 1。
求P(X < 0.5)。



3. f(x) = 1，其中0 < x < 1。
求P(X > 2)。

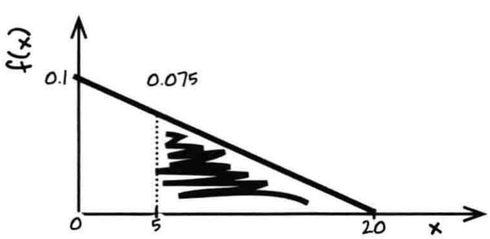
这个概率密度函数的x的上限是1，
即在大于上限时，结果为0。

P(X > 2) = 0



4. f(x) = 0.1 - 0.005x，其中0 < x < 20。
求P(X > 5)。

当x = 5时，f(x) = 0.075。即我们必须求出
高0.075，宽15的直角三角形的面积。



P(X > 5) = (0.075 × 15)/2
= 1.125/2
= 0.5625

三角形的面积等于底乘以高之积的1/2。

要点

- **离散数据**由单个数值组成。
- **连续数据**包含一个数据范围，这个范围内的任何一个数值都有可能发生。其数据常常用测量方法得到，而不是用计数方法得到。
- 连续概率分布可以用概率密度函数进行描述。
- 通过计算一个数值范围内的概率密度函数下方的面积，可得出该数值范围的概率。也就是说，为了求出$P(a < X < b)$，必须计算a和b之间的概率密度函数下方的面积。
- 概率密度函数下方的总面积必须等于1。

概率算好了

前面已经讲过如何使用概率密度函数求连续数据的概率。我们算出，朱莉为了见到约会者而需要等待5分钟以上的概率是0.75。



寻找~~灵魂~~唯一伴侣

除了青睐守时的男伴，朱莉对于她这一类女生的另一半应该有的模样也有打算。



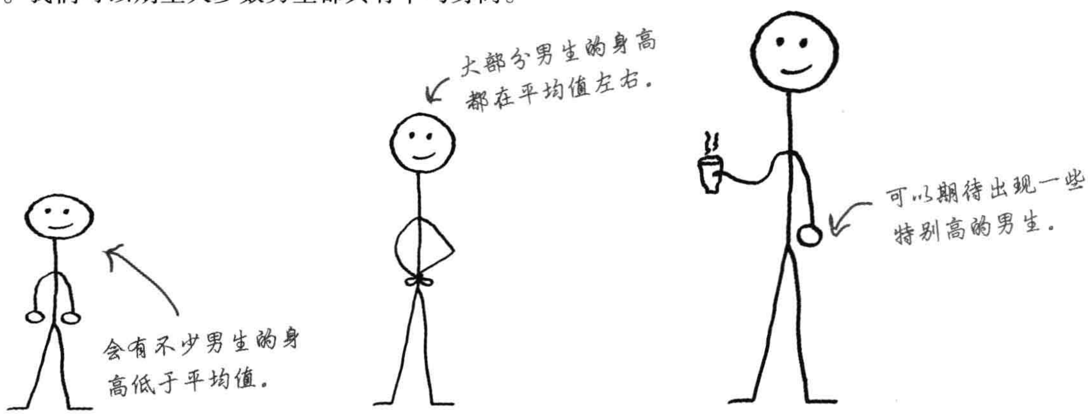
朱莉喜欢穿高跟鞋，鞋子越高她越开心。唯一的问题是，她坚持要自己的男伴在她穿最高的高跟鞋时也比她高，目前她身边没有合适的人。

可惜，前两次“相亲”的男子没有达到朱莉的预期。她想知道这些约会对象中有几个比她高，以及约会者身高够得上她的标准的概率是多少？

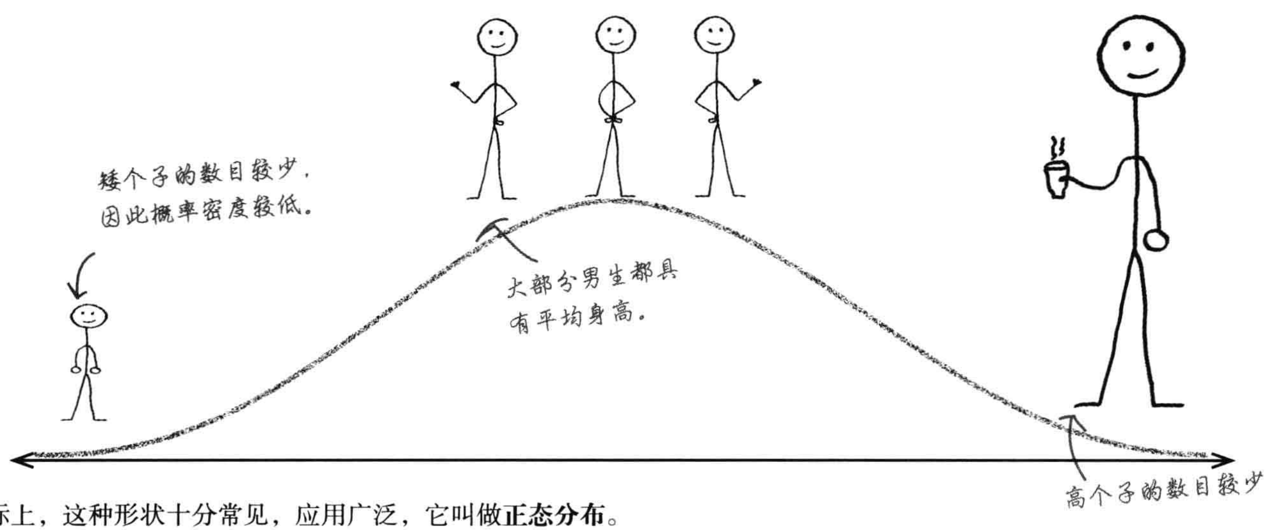
这一次我们该怎么计算概率？

男伴模型

前面讲过十分简单的连续概率分布，但那样的概率分布无法体现吸引朱莉赶赴约会的男生的身高模型。在这些男生中，很可能有几位的身高远远低于平均水平，有几位确实很高，还有很多介于以上两种情况之间。我们可以期望大多数男生都具有平均身高。



在这种给定模式下，男生身高的概率密度有可能是这个样子。



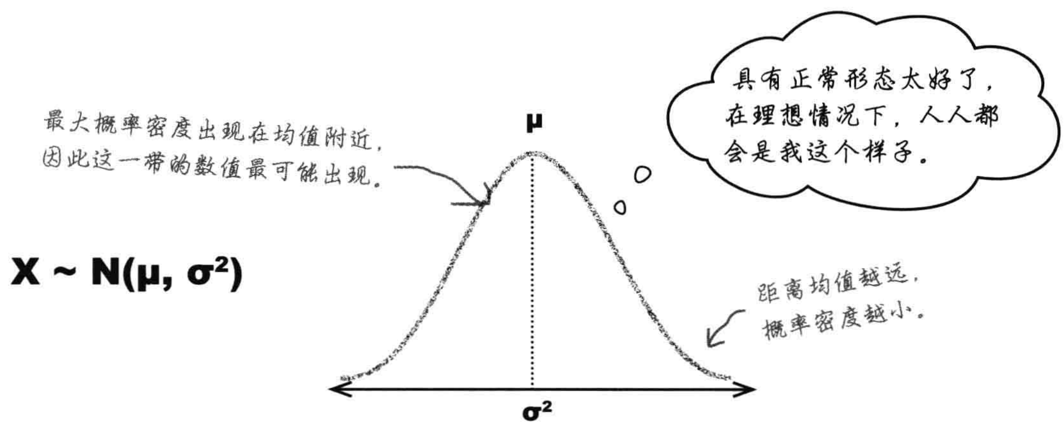
实际上，这种形状十分常见，应用广泛，它叫做**正态分布**。

正态分布是连续数据的“理想”模型

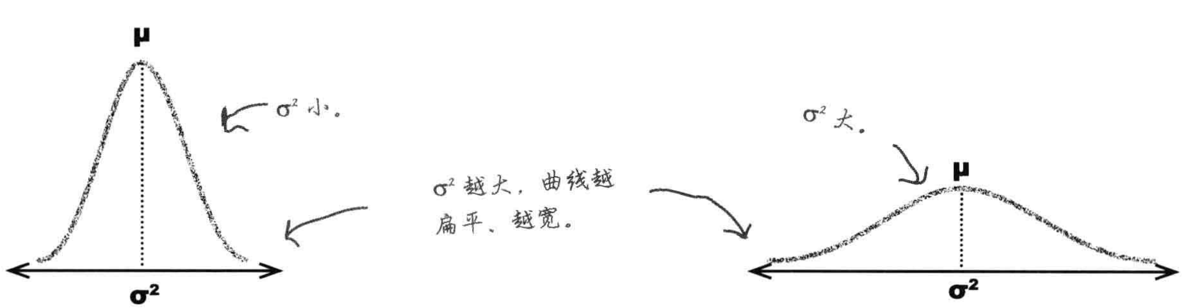
正态分布之所以被称为正态，是因为它的形态看起来合乎理想。在现实生活中，遇到测量值之类的大量连续数据时，你“正常情况下”会期望看到这种形态。

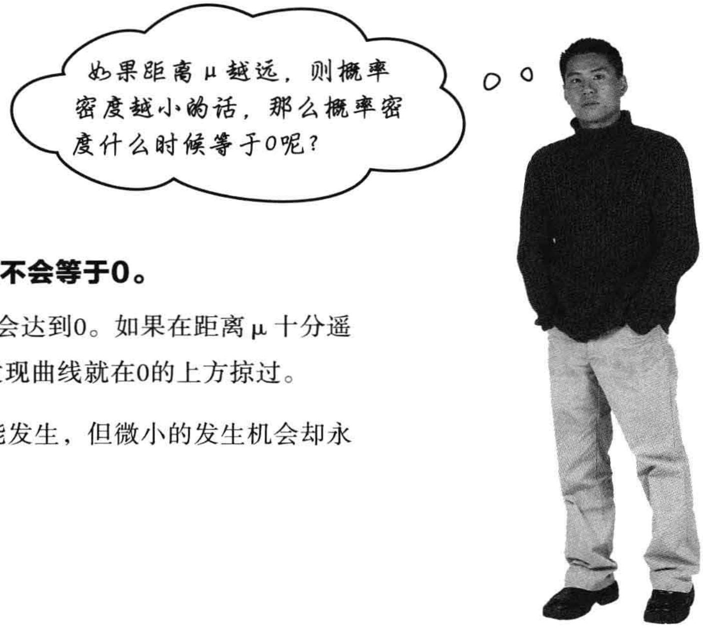
正态分布具有钟形曲线，曲线对称，中央部位的概率密度最大。越是偏离均值，概率密度减小。均值和中位数均位于中央，具有最大概率密度。

正态分布通过参数 μ 和 σ^2 进行定义。μ 指出曲线的中央位置，σ 指出分散性。如果一个连续随机变量X符合均值为 μ 、标准差为 σ 的正态分布，则通常写作 $X \sim N(\mu, \sigma^2)$。



前面讲过，μ 指出曲线的中央位置，σ^2 指出分散性。在实践中，这意味着 σ^2 越大，正态分布曲线越扁平、越宽。





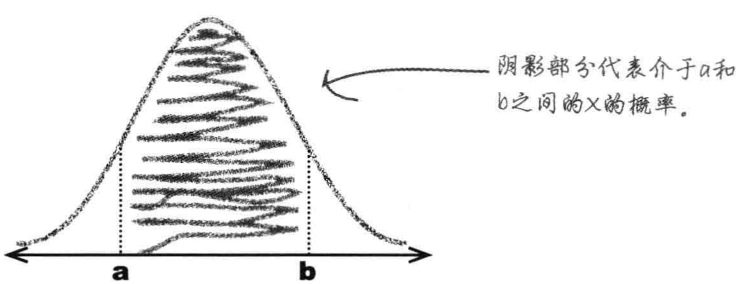
无论把图形画多大，概率密度永远不会等于0。

概率密度会越来越接近0，但永远不会达到0。如果在距离μ十分遥远的地方观察概率密度曲线，你将发现曲线就在0的上方掠过。

还可以这样理解：事件越来越不可能发生，但微小的发生机会却永远存在。

如何求正态概率？

像处理其他连续概率分布一样，可通过计算分布曲线下方的面积求出概率。曲线代表概率密度，概率则以特定范围内的面积表示。例如，如果你想求出介于a和b之间的变量X的概率，则需要求出曲线下方介于a点与b点之间的面积。



似乎很复杂？别担心，这比你想像的要容易。

如果全靠自己计算正态曲线下方的面积，难度很大。不过，幸运的是，你可以借助概率表进行查找。只要算出要求其面积的范围，再在概率表中查相应概率就行了。

正态概率计算三步法

求正态概率需分几个步骤。我们会指导你完成整个过程，不过请先看看下面这张导向图，弄清方向。

如果正态分布适用于你所遇到的情况，则看看是否能求出均值和标准差，只有先得知这些信息，才能求出概率；还需要弄清楚要求的是哪一部分面积。

❶ 确定分布与范围

现在不用担心这个步骤，很快我们就会告诉你怎么做。

❷ 使其标准化

一旦转化为正态曲线，就能使用方便易用的概率表查找概率。大功告成！

❸ 查找概率

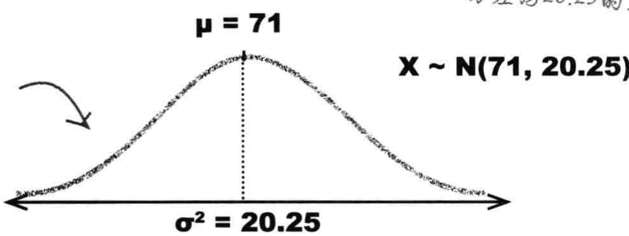
第1步：确定分布

我们要做的第一件事是确定数据分布。

朱莉已得知统计邦适龄男生的身高均值和标准差：均值71英寸，方差20.25。即，如果用X表示男生的身高，则X ~ N(71, 20.25)。

这个简明表示方法的意思是“变量X符合均值为71、方差为20.25的正态分布”。

这是X ~ N(71,20.25)的图形。

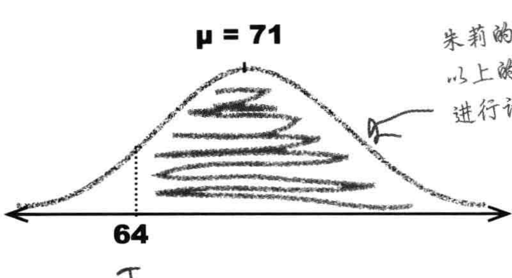


我们还需要知道哪个数值范围能得出正确的概率面积，在本例中，我们要求出与朱莉相亲的男生具有足够身高的概率。

这容易。朱莉希望她的约会者比她高，所以我们可以根据她的身高算出概率。



朱莉身高64英寸，于是我们将求出与她相亲的男生比她高的概率。



朱莉的约会者身高在64英寸以上的概率可通过这块面积进行计算：P(X > 64)。

这是X = 64英寸的位置。

第2步：标准化为N(0, 1)

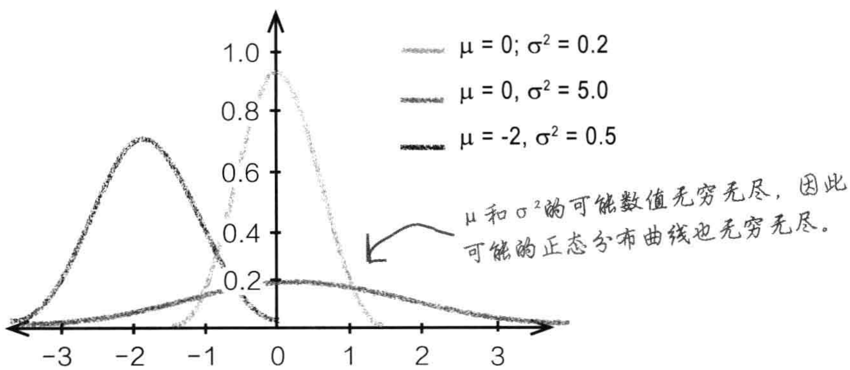
下一步是让变量X标准化，使均值为0，标准差为1，据此得出标准正态变量Z，而Z ~ N(0, 1)。





概率表仅给出N(0, 1)的概率。

概率表主要给出了N(0, 1)分布的概率，因为不可能为每一条正态分布曲线制定概率表。μ 和 σ^2的可能值无穷无尽，当正态曲线用这些数值作为参数表示曲线的中间位置和分布情况时，可能的正态分布曲线也无穷无尽。



能够利用标准正态分布意味着能够为μ 和 σ^2的所有可能数值使用同一概率表。只有一个问题：如何将正态分布转变为标准形式？

动动脑

你觉得我们可以怎样对正态分布进行标准化？

欲完成标准化，先移动均值……

让我们先进行正态分布转化，使得均值为0，而不是71，为此，将曲线向左移动71。



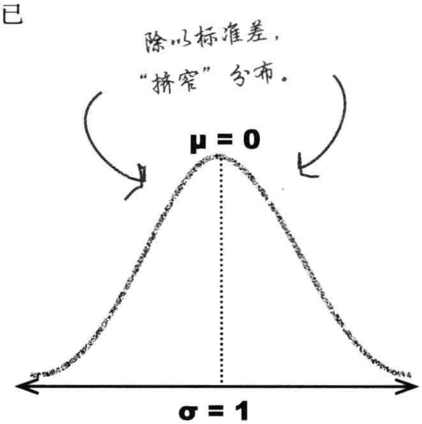
这样就得到一个新分布：

$$X - 71 \sim N(0, 20.25)$$

然后收窄

我们还需要调整方差。为此，通过除以标准差“挤窄”我们的分布。已知方差为20.25，所以标准差为4.5。

复习一下：标准差是方差的平方根。



于是得到 $\frac{X - 71}{4.5} \sim N(0, 1)$

或 $Z \sim N(0, 1)$，其中：

$$Z = \frac{X - 71}{4.5}$$

看着眼熟吗？这正是我们在第3章中首次讲到标准差时出现过的**标准分**。通常，通过下式可求出任何正态变量X的标准分：

X是我们试图求其概率的变量。

X的均值

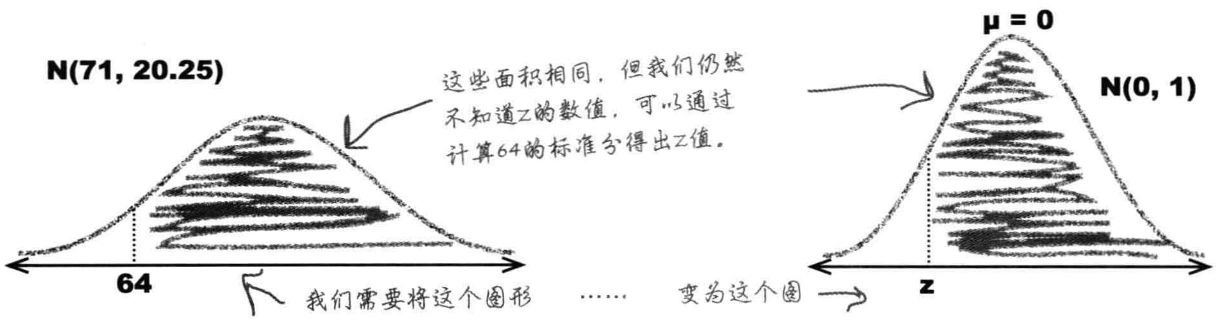
$$Z = \frac{X - \mu}{\sigma}$$

X的标准差

现在，为要计算其概率的特定数值求出Z

前面讲过如何对概率分布进行标准化，从而令$X \sim N(\mu, \sigma^2)$变为$Z \sim N(0, 1)$。我们最感兴趣的是实际概率，我们要做的是为需要求概率的数值找出数值范围，然后求出这个范围的限值的标准分，最后可以通过正态分布表查找求得标准分的概率。

在我们的例子中，需要求朱莉的约会对象比朱莉高的概率。由于朱莉的身高是64英寸，因此我们要求$P(X>64)$，这个数值范围的限值是64，所以，只要算出64的标准分z，就能据此求出概率。



让我们求出64的标准分。

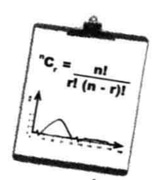
$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{64 - 71}{4.5}$$

$$= -1.56\text{(保留两位小数)}$$

所以，根据统计邦男生身高均值和标准差，算得64的标准分为-1.56。

得出这个结果后，我们就可以进入最后一步：通过概率表查找概率。



重要统计量

标准分

通过下式可求得一个数值的标准分：

$$Z = \frac{X - \mu}{\sigma}$$

世上没有傻问题

问： 这个标准分和我们以前见过的标准分是一样的吗？

答： 是一样的。正态分布不是唯一能用上标准分的地方，但是，在允许使用标准正态概率表的情况下，标准分特别有用。

问： 经过标准化的数值范围的概率的确等于原来的分布概率吗？如何实现？

答： 概率相同，而且使用概率表方便得多。

在我们对原来的正态分布进行标准化时，一切比例都保持相同。整个区间既没有增大，也没有缩小，由于代表概率的是面积，因此概率也保持不变。

动动笔

标准化时间到了。我们将给你一个分布和一个数值，请说出标准分。

1. N(10, 4)，数值：6

2. N(6.3, 9)，数值：0.3

3. N(2, 4)。如果标准分等于0.5，数值是多少？

4. 数值20的标准分是2。如果方差为16，那么均值是多少？

动动笔解答

标准化时间到了。我们将给你一个分布和一个数值，请说出标准分。

1. N(10, 4)，数值：6

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{6 - 10}{2}$$

$$= -2$$

2. N(6.3, 9)，数值：0.3

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{0.3 - 6.3}{3}$$

$$= -2$$

3. N(2, 4)。如果标准分等于0.5，数值是多少？

这是前面问题的逆运算。我们已知标准分，需要求原来的数值。通过代入已知条件可求得x。

$$z = \frac{x - \mu}{\sigma}$$

$$0.5 = \frac{x - 2}{2}$$

$$0.5 \times 2 = x - 2$$

$$x = 1 + 2$$

$$= 3$$

4. 数值20的标准分是2。如果方差为16，那么均值是多少？

这个问题与问题3相似。代入已知数值可求得μ。

$$z = \frac{x - \mu}{\sigma}$$

$$2 = \frac{20 - \mu}{4}$$

$$2 \times 4 = 20 - \mu$$

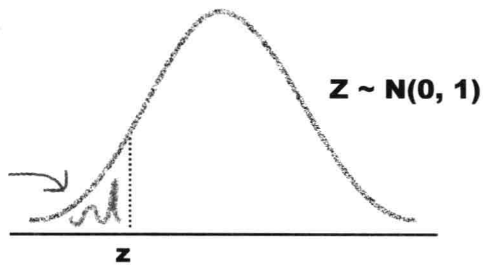
$$\mu = 20 - 8$$

$$= 12$$



第3步：用方便易用的概率表查找概率

既然已经得出了标准分，就可以用概率表求概率了。利用标准正态概率表可以查找任何z值，进而查出相应概率P(Z < z)。





放轻松

我们已将需要使用的各种概率表放在附录II中。

翻到658-659页，利用正态分布表查找本章要求计算的概率。

如何使用概率表？

先算z，保留两位小数，这就是你要在表中查找的数值。

查找概率时，需要用第一列和第一行找出数值z，第一列为z值（保留一位小数，不进行四舍五入），第一行为第二位小数，两行的交点即为概率。

例如，如想求P(Z < -3.27)，则在第一列找到-3.2，在第一行找到.07，然后找出概率0.005。

这一行代表.07，即z的第二位小数。

这一行代表z = -3.2x，其中x是某个数字。

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064

这是-3.2和.07的交点，等于P(Z < z)的数值。

朱莉要算的概率就在表中

让我们回头看朱莉的问题，我们需要求P(Z > -1.56)，因此，让我们在概率表中查找-1.56，看看结果如何。

在本书末尾的附录部分可找到正态概率表。

这是代表z的第二位小数0.06的列。

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170

这是代表z = -1.5x的行，其中x是某个数值。

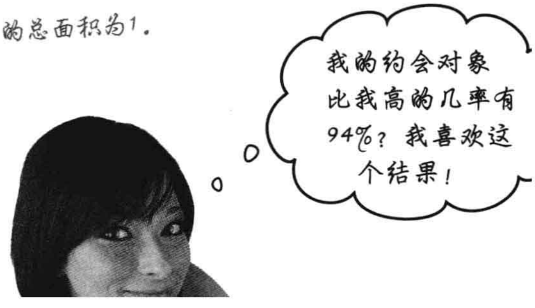
这是-1.5和0.06的交点，这就是P(Z < z)的数值。

结果，在概率表中查找-1.56，得出概率0.0594，即P(Z < -1.56) = 0.0594，这表示：

$$
\begin{aligned}
P(Z > -1.56) &= 1 - P(Z < -1.56) \\
&= 1 - 0.0594 \\
&= 0.9406
\end{aligned}
$$

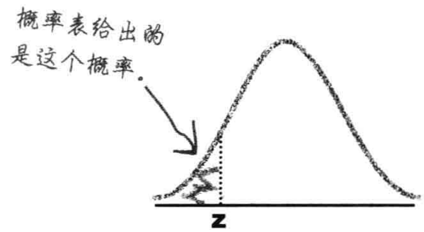
总概率为1，因此曲线下的总面积为1。

也就是说，朱莉的约会对象比她高的概率是0.9406。



概率表细细看

通过概率表可查找$P(Z<z)$的概率，其中z为某个数值。问题来了：你要求的并不总是这一类概率；有时候你需要求一个大于z的连续随机变量的概率，或是介于某两个数值之间的一个连续随机变量的概率。这时如何通过概率表求出所需要的概率？



为了利用概率表求出需要的结果，需要好好动动脑筋，通常的做法是求出一个整体面积，然后减去不需要的部分。

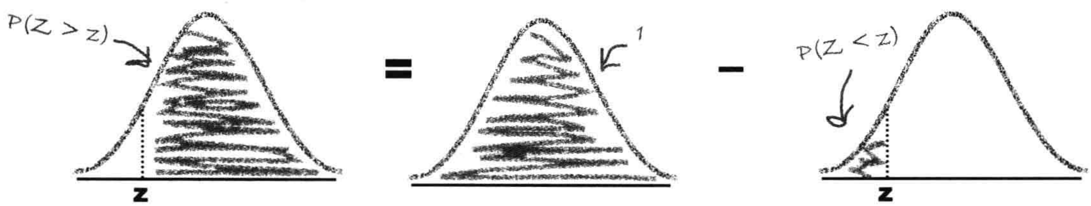
求解$P(Z > z)$

$P(Z > z)$类型的概率可通过以下方法求解：

$$P(Z > z) = 1 - P(Z < z)$$

我们已经利用这个式子求出了朱莉比约会者高的概率。

即，将$Z < z$的面积从总概率中去除。



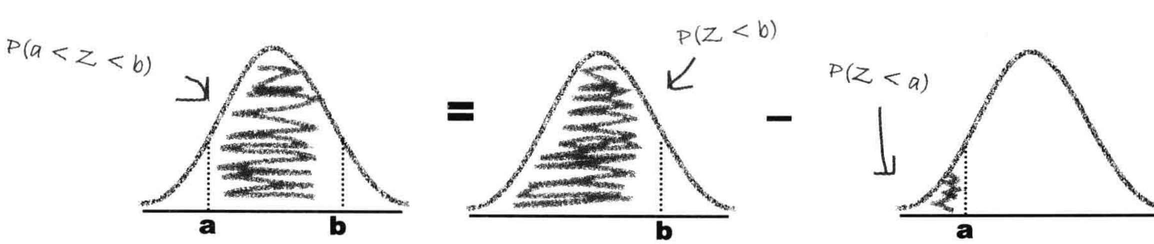
求解$P(a < Z < b)$

这一类概率的算法略微复杂一点儿，但仍然能够得到解答。可通过下列算法进行计算：

$$P(a < Z < b) = P(Z < b) - P(Z < a)$$

用这个式子可以算出朱莉的约会者的身高在某个特定范围以内的概率。

即，算出$P(Z < b)$，然后将$P(Z < a)$面积从其中去除。



世上没有傻问题

问： 我曾经听说过“高斯”这个术语，它指的是什么？

答： 正态分布的另一个名称是高斯分布。如果你听见别人在谈论高斯分布，那么他们就是在谈论正态分布。

问： 所有的正态概率表都相同吗？

答： 所有的正态概率表都能给出相同的概率。不过，概率表的实际覆盖范围会有一些变化。

问： 变化？什么意思？

答： 有的制表和考试委员会为概率表设定不同的精度等级，还有一些会以略有不同的格式制作表格，但表中的信息都是一样的。

问： 如果我要参加概率考试该怎么办？

答： 首先了解考试中使用的概率表的格式，然后看看能不能搞一份复印件。

得到考试委员会采用的概率表后，花点时间熟悉熟悉，这样你就能在考试到来时轻易过关了。

问： 求一个范围的概率似乎有些棘手，我该怎么做？

答： 关键在于想办法通过概率表求出要求的面积。概率表通常只给出$P(Z<z)$形式的概率，其中z为某个数值。因此，最大的困难就在于把你要求的概率改写成符合这种形式的概率。

如果所计算的是$P(a<Z<b)$形式的概率，即某个范围的概率，则需要查找两个概率，一个是$P(Z<a)$的概率，另一个是$P(Z<b)$的概率，查到这两个概率后，用最大的概率减去最小的概率就行了。

问： 连续分布有众数吗？你能求出正态分布的众数吗？

答： 有。连续概率分布的众数即概率密度最大处的数值。如果画出概率密度，则众数为曲线最高点处的数值。

观察正态分布曲线，可以看到最高点位于正中央。正态分布的众数为μ。

问： 中位数呢？

答： 一个连续概率分布的中位数即$P(X<a)=0.5$处的数值，即将概率密度曲线下方的面积一分为二的数值。

正态分布的中位数也是μ。在处理连续概率分布时，中位数和众数并不那么常用，期望和方差更为重要。

问： 什么是标准分？

答： 一个变量的标准分即用这个变量减去其均值再除以这个变量的标准差的商。这是对正态分布进行标准化的一种方法，可令正态分布转化为$N(0,1)$分布，从而可以对各种正态分布进行比较。在处理正态分布时，标准分很有用，因为这样一来，你就可以通过标准正态概率表查找概率。

一个特定数值的标准分还说明了数值与均值相距多少个标准差，你可以由此获悉该数值与均值的相对接近程度。

动动笔

现在该考考你的概率表使用技术了，看看是否能解答以下概率问题。

1. $P(Z < 1.42)$。

2. $P(-0.15 < Z < 0.5)$。

3. $P(Z > z) = 0.1423$。z等于多少？

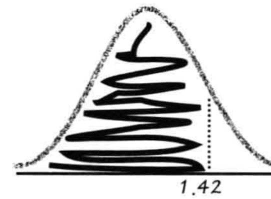
动动笔解答

现在该考考你的概率表使用技术了，看看是否能解答以下概率问题。

1. P(Z < 1.42)。

在概率表中查1.42可以求出这个概率，结果为：

$$P(Z < 1.42) = 0.9222$$



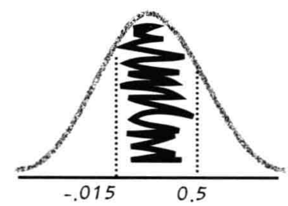
2. P(−0.15 < Z < 0.5)。

查找P(Z < 0.5)，然后减去P(Z < −0.15)

$$P(-0.15 < Z < 0.5) = P(Z < 0.5) - P(Z < -0.15)$$

$$= 0.6915 - 0.4404$$

$$= 0.2511$$



3. P(Z > z) = 0.1423。z等于多少?

这个问题略有难度：已知概率，要求z值。

已知P(Z > z) = 0.1423，即：

$$P(Z < z) = 1 - 0.1423$$

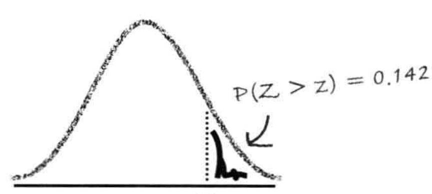
$$= 0.8577$$

接下来要求出哪个z值的概率为0.8577，通过概率表查出：

$$z = 1.07$$

所以

$$P(Z > 1.07) = 0.1423$$



练习



朱莉有一个问题，当我们计算她的约会对象比她个子高的概率时，没有把她的高跟鞋算上。看看你能不能求出朱莉穿上5英寸高的高跟鞋时，她的约会者比她高的概率？

提醒一下，朱莉身高64英寸，X ~ N(71, 20.25)，X为统计邦男生的身高。



朱莉有一个问题，当我们计算她的约会者比她个子高的概率时，没有把她的高跟鞋算上。看看你能不能求出朱莉穿上5英寸高的高跟鞋时，她的约会者比她高的概率？

提醒一下，朱莉身高64英寸，X ~ N(71, 20.25)，X为统计邦男生的身高。

当朱莉穿上5英寸高的高跟鞋后，她的身高变为69英寸。我们需要求P(X > 69)。

我们需要先求178的标准分，这样才能用概率表查找相应概率。

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{69 - 71}{4.5}$$

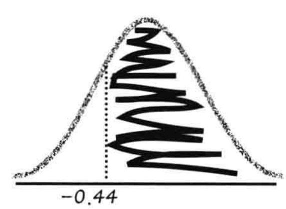
← 方差为20.25，因此其平方根，也就是标准差，为4.5。

$$= \frac{-2}{4.5}$$

$= -0.44$（保留两位小数）

现在我们已经求出了z，需要继续求P(Z > z)，即P(Z > −0.44)。

$$P(Z > -0.44) = 1 - P(Z < -0.44)$$
$$= 1 - 0.3300$$
$$= 0.67$$



因此，在朱莉穿上5英寸高的高跟鞋后，她的约会对象比她高的概率是0.67。



案件：缺失的参数

维尔在芒芒游戏公司工作，他遇到了一个问题。他需要向老板报告人们闯过新游戏第一关所花时间(分钟)的均值和标准差。这倒不难，可不巧的是，一头恶犬咬掉了他写有概率的那张纸。



威尔只有3条有用线索。

首先，威尔知道人们闯过第一关所用的时间符合正态分布。

其次，他知道一位玩家的闯关时间少于5分钟的概率为0.0045。

最后，某个人闯过第一关花费的时间少于15分钟的概率是0.9641。

威尔如何求出均值和标准差?

破案：缺失的参数

威尔如何求出均值和标准差？



威尔可以使用概率表和标准分得出均值和标准差的表达式，然后求解。

首先，我们知道$P(X<5)=0.0045$，从概率表上看，$P(X<z_1)$，其中$z_1=-2.61$，即5的标准分为-2.61。如果将这个结果代入标准分公式，得到：

$$-2.61=\frac{5-\mu}{\sigma}$$

类似地，$P(X<15)=0.9641$，即15的标准分等于1.8，我们得到：

$$1.8=\frac{15-\mu}{\sigma}$$

这样我们就得到两个等式，可以求解μ和σ。

$$\left.\begin{aligned}-2.61\sigma &= 5-\mu\\ 1.8\sigma &= 15-\mu\end{aligned}\right\}$$

我们现在可以解这个方程组。

用第二个等式减去第一个等式，得：

$$\begin{aligned}1.8\sigma+2.61\sigma &= 15-\mu-5+\mu\\ 4.41\sigma &= 10\\ \sigma &= 2.27\end{aligned}$$

将以上结果代入第二个方程，得：

$$\begin{aligned}1.8\times 2.27 &= 15-\mu\\ \mu &= 15-4.086\\ &= 10.914\end{aligned}$$

即：

$$\left.\begin{aligned}\mu &= 10.914\\ \sigma &= 2.27\end{aligned}\right\}$$

这就是μ和σ的值。

从那以后，他们幸福地生活在一起

概率算得很准，朱莉在上一次“相亲”中成功了！为了保证未来的灵魂伴侣能和她的鞋子般配，朱莉挑出最高的高跟鞋穿上，对他进行测试。还有，当她来到约会地点的时候，他已经在那儿了，她不用等呢。



可事情尚未到此为止。

继续看书吧，我们将向你介绍更多有关正态分布的知识，目前你不过是触及皮毛哦。

要 点

- 数据由单个数值组成。正态分布的形状为对称的钟形，其定义为$N(\mu, \sigma^2)$。
- 求正态概率时，首先要确定所需要的概率范围，然后求出这个范围的限值的标准分，算式如下：

$$Z = \frac{X - \mu}{\sigma} \quad 其中Z \sim N(0, 1)。$$

- 通过在概率表中查找标准分可求出正态概率，概率表给出的是等于或者小于这个数值的概率。

9 再谈正态分布的运用

超越正态



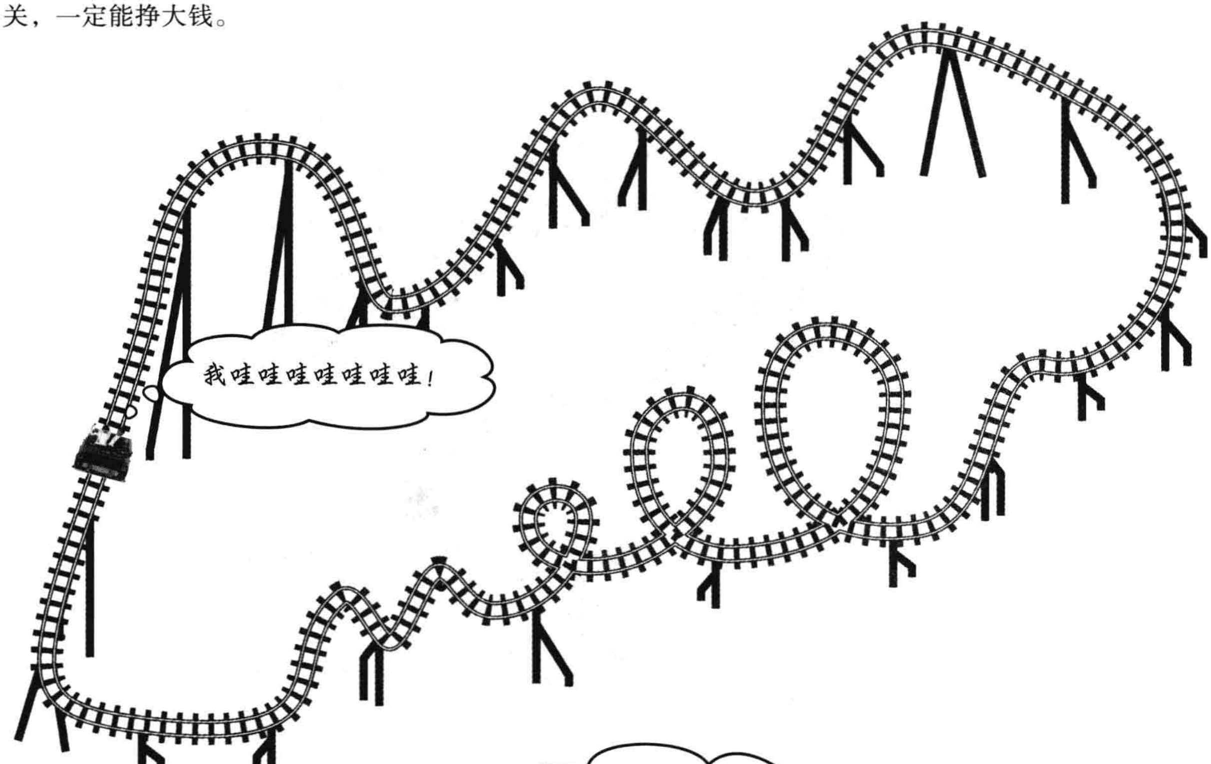
但愿所有的概率分布都是正态分布。

有了正态分布，日子**好过多了** — 既能一口气查出整个范围的概率，又能留下点时间玩游戏，谁还会花时间一个一个地计算概率呢？在本章中，你将学习如何闪电般**解决更复杂的问题**，还将懂得如何将正态分布的便利运用到**其他概率分布**上。

爱情就像过山车

如今婚礼筹办市场生意红火，为了让顾客对这个特别的日子刻骨铭心，德克想出了一个好主意。干嘛一定要在地面上办婚礼呢？坐过山车不是更好吗？

德克对这个“爱情过山车”创意很有信心，认为只要能过健康和安全这一关，一定能挣大钱。







在大刀阔斧开展业务之前，德克需要确保他所设想的特别座驾能够承载新郎和新娘的重量，所以请你看看能不能帮个忙。

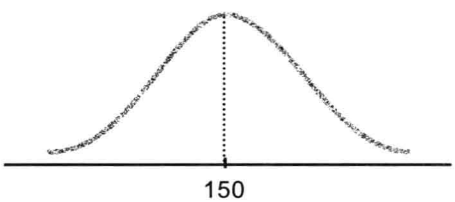
他所设想的座驾能够承载最多380磅的重量。新郎和新娘综合体重不超过这个重量的概率是多少？

双双登上爱情过山车

在开始计算之前，我们需要了解统计邦新郎新娘的体重分布情况——包括结婚礼服在内。新郎和新娘的体重都符合正态分布，新娘的体重符合N(150, 400)，新郎的体重符合N(190, 500)，体重单位为“磅”。

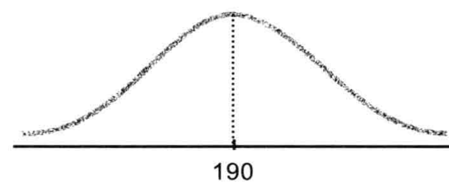
新娘 ~ N(150，400)





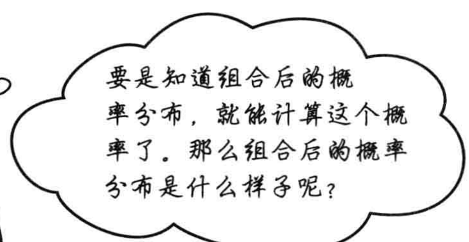
新郎 ~ N(190，500)





我们需要设法通过这两个概率分布算出一对新郎新娘的体重低于过山车允许的最大载荷的概率。如果算出的概率足够高，我们就可以满怀信心地说：坐过山车举行婚礼的想法是可行的。





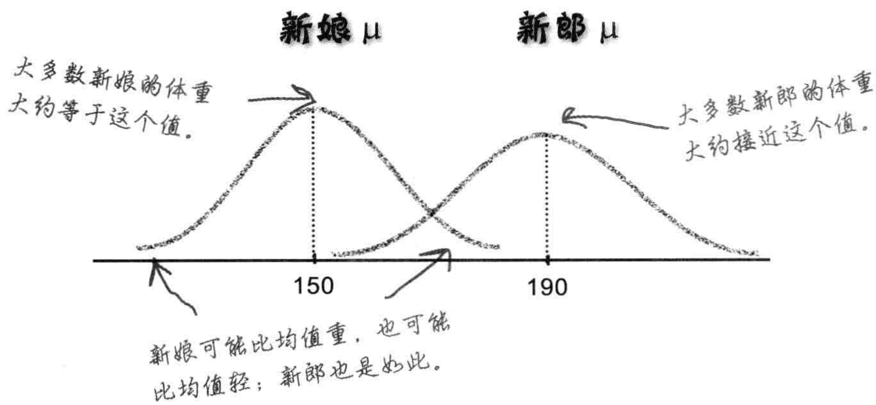
动动脑

你觉得我们该怎样求出新郎新娘综合体重的概率分布？你觉得会是哪种概率分布？为什么？

正态新娘+正态新郎

让我们先仔细看看新郎和新娘的体重分布情况。

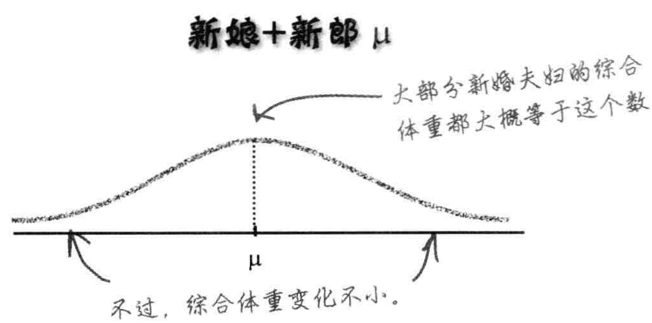
如你所知，新郎和新娘的体重符合正态分布，如下所示：



不过，我们真正要求的却是新郎和新娘的综合概率分布，即，要求新郎与新娘体重之和的概率分布。

新娘体重+新郎体重~？

假设新娘和新郎的体重互相独立，则分布形状应与下图有几分相似：



终究还是体重问题

还记得我们最开始讲到连续数据的时候吗？那时我们讲过身高、体重之类的数据往往符合什么分布来着？—— 我们那时讲到，身高、体重之类的数据是连续数据，且往往符合正态分布。

这一次，我们研究的是一对新婚佳偶的综合体重。综合体重也是体重，同时我们已经知道体重的分布趋势；综合体重依然是**连续数据**，而且，综合体重依然**符合正态分布**。这就是说，新娘加新郎的体重符合正态分布。

新娘加新郎的综合体重符合正态分布这个结论对我们大有用处。这说明我们可以像前面一样，利用概率表查找概率，即，我们可以查出综合体重低于380磅的概率 —— 这是爱情过山车的要求。

只有一个问题 —— 在动手查找概率之前，我们需要知道新娘新郎综合体重的均值和方差。该怎么求呢？

新娘和新郎的综合体重符合正态分布，但均值和方差是多少呢？

新娘 + 新郎~ N(?, ?)

动动笔

现在考考你的记忆力。还记得下列公式的简捷算法吗？假定X和Y是独立变量。

1. E(X + Y)

2. Var(X + Y)

3. E(X - Y)

4. Var(X - Y)

动动笔解答

现在考考你的记忆力。还记得下列公式的简捷算法吗？假定X和Y是独立变量。

1. E(X + Y)

   $E(X + Y) = E(X) + E(Y)$

2. Var(X + Y)

   $Var(X + Y) = Var(X) + Var(Y)$

3. E(X - Y)

   $E(X - Y) = E(X) - E(Y)$

4. Var(X - Y)

   $Var(X - Y) = Var(X) + Var(Y)$

记得吗？即使是X－Y，我们计算方差时也用的是加法。



这些简捷算法也适用于连续数据。

我们最初讲到这些简捷算法的时候，用的是离散数据。幸运的是，同样的计算规则和简捷算法也适用于连续数据。

动动脑

你认为我们该怎样用这些简捷算法求出新郎新娘体重之和的概率分布？

综合体重符合哪种分布？

前面已经讲过，新郎新娘的综合体重符合正态分布，这说明我们可以利用概率表查找综合体重低于某个特定值的概率。

让我们试试用X和Y表示新郎新娘的体重分布，如果用X代表新娘的体重，用Y代表新郎的体重，则X和Y是独立的，然后需要求出μ和σ，其中：

$$X + Y \sim N(\mu, \sigma^2)$$

X＋Y表示“新娘的体重＋新郎的体重”，我们如何获知它们的概率分布均值和方差呢？

也就是说，在进一步进行计算之前，我们需要求出X + Y的期望和方差，怎么求？

查看前一个练习的答案，可以看出，当我们处理离散概率分布时，只要X和Y是独立变量，就可以用下列算式计算E(X + Y)和Var(X + Y)：

$$E(X + Y) = E(X) + E(Y) \quad 且 \quad Var(X + Y) = Var(X) + Var(Y)$$

于是，只要知道X和Y的期望和方差，就能用上面的式子计算X+Y的期望和方差。



也就是说，只要我们知道X和Y的概率分布，就能同时算出X+Y的概率分布。

我们可以用已知求未知。

由于我们已知新娘体重和新郎体重的概率分布，因此能求出新郎新娘综合体重的概率分布。

让我们仔细看看。

X + Y概率分布细细看

在研究综合正态变量的时候，想办法求出X+Y的分布是十分有用的。如果独立随机变量X和Y符合正态分布，那么X+Y也符合正态分布。另外，你还可以使用X和Y的均值和方差计算X+Y的概率分布。

记住：如果两个变量互相对对方的概率没有影响，则这两个变量是相互独立的。

为了求出X+Y的均值和方差，可以使用离散概率分布的相同计算公式，即，如果：

$$X \sim N(\mu_x, \sigma_x^2) \quad 且 \quad Y \sim N(\mu_y, \sigma_y^2)$$

则

$$\mathbf{X + Y \sim N(\mu, \sigma^2)}$$

将X和Y的均值相加可得到X+Y的均值，类似地，将X和Y的方差相加可得到X+Y的方差。

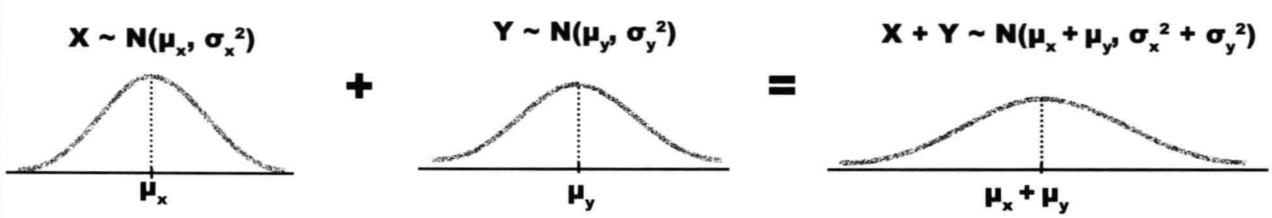
其中

$$\mu = \mu_x + \mu_y \qquad \sigma^2 = \sigma_x^2 + \sigma_y^2$$

如果X和Y是独立变量，则可以使用这些简捷算法——这样日子就好过多了。

即，X + Y的均值等于X的均值加上Y的均值，X + Y的方差等于X的方差加上Y的方差。

查看以下草图，注意到X + Y的方差的特点了吗？



X + Y的方差大于X的方差，也大于Y的方差，这使得X + Y的曲线比X的曲线和Y的曲线都拉得长，这一点对于任何正态X和Y都成立。在将两个变量相加之后，实际上增大了变异性，于是使得分布形状拉长；随着图形拉长，图形还会变得更扁，这样才能使图形下方的总面积仍然为1。

X – Y概率分布细细看

有时候，X + Y并不是你要求的概率，如果所求的是两个变量之差的概率，则需要计算X – Y。

如果X和Y是独立随机变量，且都符合正态分布，则X – Y符合正态分布，这一点和X + Y的规律完全一样。

为了求出均值和方差，我们再次使用离散概率分布的同一组简捷算法，只要：

$$X \sim N(\mu_x, \sigma_x^2)且Y \sim N(\mu_y, \sigma_y^2)$$

则

$$\mathbf{X - Y \sim N(\mu, \sigma^2)}$$

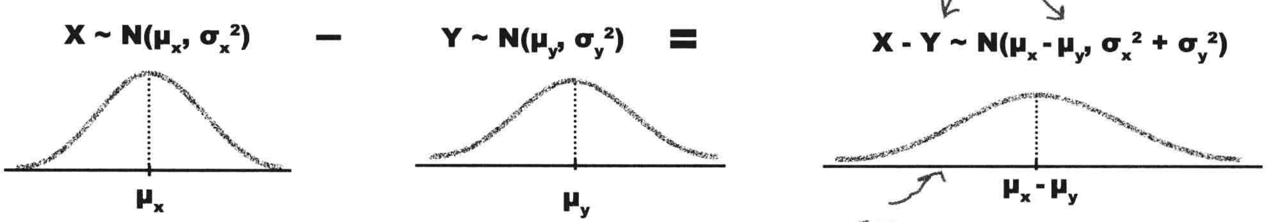
方差用加法计算，这和离散概率分布的计算一模一样。

其中

$$\mu = \mu_x - \mu_y \qquad \sigma^2 = \sigma_x^2 + \sigma_y^2$$

即，X – Y的均值等于X的均值减去Y的均值，X – Y的方差等于X的方差加上Y的方差。

均值用减，方差用加。



看这图形的形状，和X + Y的形状一样，这是因为方差相同；但曲线的中心位置变了。

方差的加法计算一眼看上去并不直观，不过，这和计算离散概率分布的道理是一样的，尽管我们用X减去Y，但实际上变异性还是增大了，方差之和反映了这种变化。和X+Y的分布一样，无论是与X相比还是与Y相比，X-Y都导致图形拉长、变扁。

查看X – Y的形状，可以看出该曲线形状和X + Y的曲线形状一样，只不过中心位置发生了移动。两种概率分布的方差相同，均值各异。

求解概率

既然知道如何计算X + Y的概率分布，就让我们看看如何利用这个概率分布计算概率。步骤如下：

我们知道，我们需要利用X + Y，而且有办法算出均值和方差。

❶ 算出分布和范围

知道分布和范围后，即可进行标准化。

❷ 将分布标准化

随后可在标准正态概率表中查找概率。

❸ 查找概率

感觉似曾相似？这些步骤和上一章中的正态分布的计算步骤是一模一样的。

世上没有傻问题

问： 告诉我，为什么我们需要求X + Y的分布？

答： 我们所求的是新郎新娘综合体重低于380磅的概率，即需要知道综合体重的分布情况。我们用X代表新娘的体重，用Y代表新郎的体重，因此需要求X + Y的分布。

问： 你说我们可以用概率表查X + Y的概率。怎么做呢？

答： 和以前的做法一模一样：找出概率分布，算出标准分，然后在概率表中查找。查找X+Y的概率和查找别的变量的概率并无区别，只要求出标准分，即可查找出所求概率。

问： 这么说，我们用来计算离散数据的简捷算法同样适用于连续数据？

答： 不错，是这样。这样就可以方便地将随机变量综合起来，求出其分布方式，进而解答更复杂的问题。

关键要记住，只有在变量为独立变量时，这些简捷算法才适用。

问： 能告诉我“独立”是什么意思吗？

答： 如果两个变量互为独立变量，则它们相互之间对对方的概率没有影响。在我们所举的例子中，我们假定新娘的体重不受新郎的体重的影响。

问： 如果X和Y不独立呢？情况会如何？

答： 如果X和Y不独立，则我们无法使用这些简捷算法，而需要大动干戈地求出X+Y的分布，这样才能得出X和Y之间的关系。

动动笔

通过下列3个步骤求出新娘和新郎的综合体重少于380磅的概率。

1. X为新娘体重，Y为新郎体重，且$X \sim N(150, 400)$，$Y \sim N(190, 500)$。根据以上信息，求出新郎新娘综合体重的概率分布。

2. 然后，利用所求出的概率分布，计算380磅的标准分。

3. 最后，利用标准分查出$P(X + Y < 380)$。

动动笔解答

通过下列3个步骤求出新娘和新郎的综合体重少于380磅的概率。

1. X为新娘体重，Y为新郎体重，且$X \sim N(150, 400)$，$Y \sim N(190, 500)$。根据以上信息，求出新郎新娘综合体重的概率分布。

我们需要求X+Y的概率分布，为了求出X+Y的均值和方差，我们将X和Y各自的均值和方差加起来，得到：

$$X + Y \sim N(340, 900)$$

2. 然后，利用所求出的概率分布，计算380磅的标准分。

$$z = \frac{(x + y) - \mu}{\sigma}$$

$$= \frac{380 - 340}{30}$$

$$= \frac{40}{30}$$

$= 1.33$（保留两位小数）

还记得我们以前用过的$z = \frac{x - \mu}{\sigma}$吗？

这一次，我们用的是$X + Y$的概率分布，

因此$z = \frac{(x + y) - \mu}{\sigma}$

3. 最后，利用标准分查出$P(X + Y < 380)$。

如果我们在标准正态概率表中查找1.33，得到概率0.9082，即：

$$P(X + Y < 380) = 0.9082$$



练习

朱莉的媒人又忙开了。一名男子至少比一名女子高5英寸的概率是多少？

在统计邦，身高以英寸计量，男性身高的概率分布为N(71,20.25)，女性身高的概率分布为N(64,16)。

练习解答

朱莉的媒人又忙开了。一名男子至少比一名女子高5英寸的概率是多少？

在统计邦，身高以英寸计量，男性身高的概率分布为N(71,20.25)，女性身高的概率分布为N(64,16)。

让我们用X代表男性身高，用Y代表女性身高，即：$X \sim N(71, 20.25)$，$Y \sim N(64, 16)$。

我们需要求出一名男子比一名女子至少高5英寸的概率，即需要求：

$$P(X > Y + 5)$$

或

$$P(X - Y > 5)$$

为了求出X－Y的均值和方差，我们用X的均值减去Y的均值，得到：

$$X - Y \sim N(7, 36.25)$$

我们需要求出5英寸的标准分：

$$z = \frac{(x - y) - \mu}{\sigma}$$

$$= \frac{5 - 7}{6.02}$$

$$= -0.33 \text{（保留两位小数）}$$

于是可以求出$P(X - Y > 5)$.

$$P(X - Y > 5) = 1 - P(X - Y < 5)$$

$$= 1 - 0.3707$$

$$= 0.6293$$

更多人想坐爱情过山车

看来，新郎新娘的综合体重小于过山车限额载荷的几率很大，不过，为什么仅限新郎新娘乘坐过山车呢？

客户们要求让更多婚礼宾客登上过山车，他们愿意出大价钱。这太好了，不过，爱情过山车承受得了这些额外的负载吗？



让我们再加上一辆轿车，另外载上四位婚礼成员，看看结果如何。这些成员可能会是老爸、老妈、伴娘、伴郎或新娘新郎希望共同登车的任何人。

轿车的总载重量为800磅，假定一位成年人的体重分布为：

$$X \sim N(180, 625)$$

其中X代表一位成年人的体重，单位为“磅”。可是如何计算4位成年人的综合体重低于800磅的概率呢？

动动脑

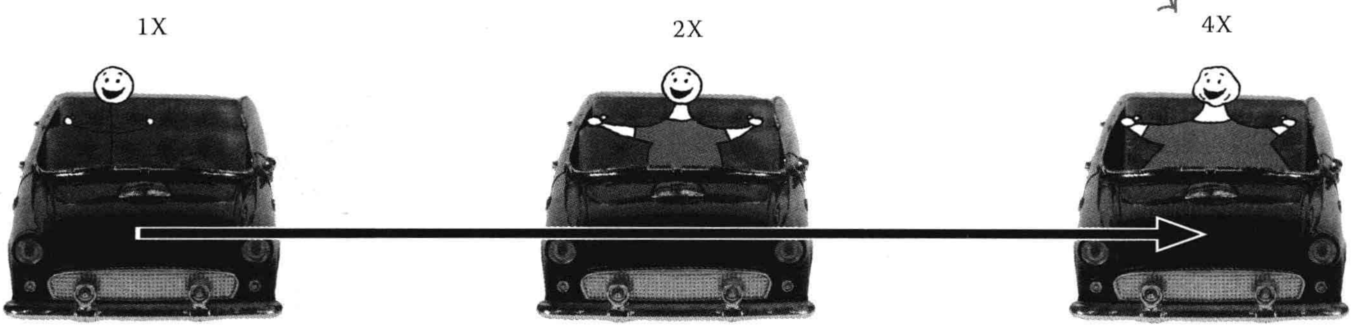
回头想想计算期望和方差时用过的简捷算法，独立观察结果和线性变换之间有何差别？二者分别对期望和方差有何影响？哪一种算法更适合解决这里的问题？

线性变换描述了数据的基本变化……

让我们先看4X的概率分布，其中X为一位成年人的体重。4X是否适合描述4位成年人的概率分布？

4X的分布其实是X的一个线性变换，是X进行aX + b变换的结果，其中a等于4，b等于0，这与我们先前在离散概率分布中遇到过的变换类型完全相同。

线性变换描述的是概率分布中的数值在大小方面的基本变化，即，4X其实描述的是一个成年人的体重放大四倍后的结果。



那么线性变换的分布是怎样的？

假定你有一个X的线性变换，其形式为aX+b，其中$X \sim N(\mu, \sigma^2)$，由于X符合正态分布，于是aX + b也属于正态分布。但期望和方差是多少呢？

让我们先算期望。在讲离散概率分布的时候，我们发现$E(aX + b) = aE(X) + b$。现在，X符合正态分布且$E(X) = \mu$，于是我们得出$E(aX + b) = a\mu + b$。

方差的处理方法与此相似，在讲离散概率分布的时候。我们发现$Var(aX + b) = a^2 Var(X)$，且这里的$Var(X) = \sigma^2$，于是得出$Var(aX + b) = a^2\sigma^2$。

合并以上两个结果，得到：

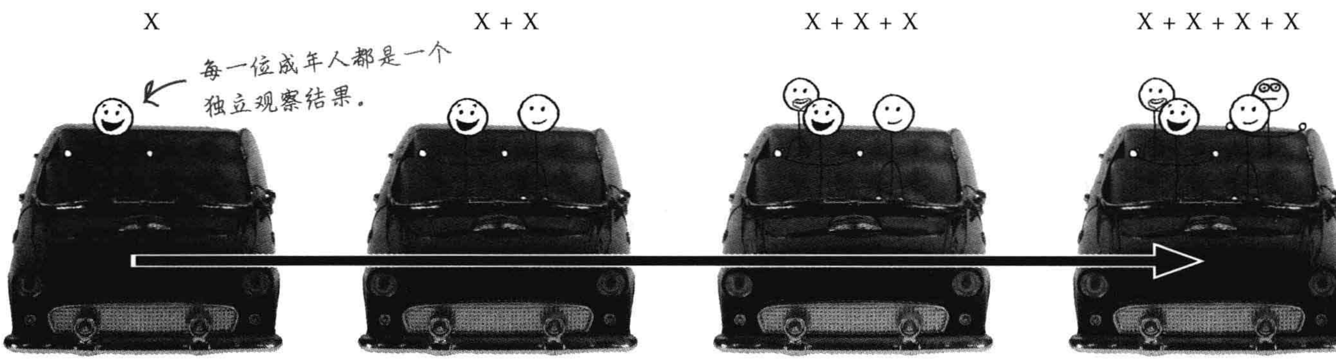
新方差是a的平方与原方差的乘积。

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

即，新均值为$a\mu + b$，新方差为$a^2\sigma^2$。那么独立观察结果是多少？

而独立观察结果描述的是你有多少数值

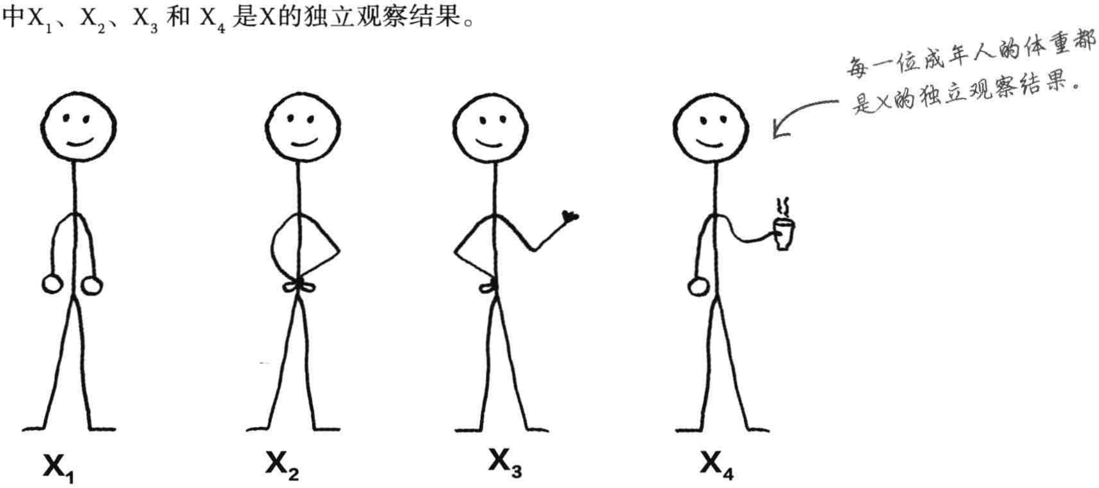
我们实际需要计算的是4位独立成年人的综合体重的概率分布，而不是对每一位成年人的体重进行变换。即，我们需要算出X的4个独立观察结果的概率。



每一位成年人的体重都是X的一个观察结果，这意味着每一位成年人的体重都通过X的概率分布进行描述。我们需要算出X的4个独立观察结果的概率分布，也就是要求以下概率：

$$X_1 + X_2 + X_3 + X_4$$

其中X_1、X_2、X_3和 X_4是X的独立观察结果。



独立观察结果的期望和方差

在讲到离散随机变量的独立观察结果的方差和期望时，我们曾经发现：

$$E(X_1 + X_2 + \cdots X_n) = nE(X)$$

及

$$Var(X_1 + X_2 + \cdots + X_n) = nVar(X)$$

如你所料，相同的算法也适用于连续随机变量，即，如果$X \sim N(\mu, \sigma^2)$，则：

$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$$

世上没有傻问题

问： 线性变换和独立观察结果之间有何差别？

答： 线性变换影响概率分布中的基本数值。例如，如果你有一根特定长度的绳子，那么，进行线性变换会影响绳子的长度。

独立观察结果影响所处理的事件的数量。例如，如果一段绳子有n个独立观察结果，则所讨论的就是n段绳子。

通常，如果数量发生变化，则所面对的是独立变量；如果基本数据发生变化，则所面对的是变换。

问： 我真的需要分清楚哪是哪吗？这有什么区别？

答： 你必须分清楚哪是哪，因为这会影响概率计算。对于线性变换和独立观察结果，均值的计算方法是相同的，但方差的计算方法有很大差别。如果存在n个独立观察结果，则新方差是原方差的n倍。如果将概率分布按照aX + b的形式进行线性变换，则新方差为原方差的a^2倍。

问： 我能在同一个概率分布中既拥有独立观察结果又拥有线性变换吗？

答： 可以。在计算概率分布的时候，只要遵守方差和期望的基本计算规律即可。离散概率分布和连续概率分布的规律是相同的。

要点

- 如果$X \sim N(\mu_x, \sigma^2_x)$，$Y \sim N(\mu_y, \sigma^2_y)$，且X和Y为独立变量，则：

  $X + Y \sim N(\mu_x + \mu_y, \sigma^2_x + \sigma^2_y)$

  $X - Y \sim N(\mu_x - \mu_y, \sigma^2_x + \sigma^2_y)$

- 如果$X \sim N(\mu, \sigma^2)$且a和b都是数字，则：

  $aX + b \sim N(a\mu + b, a^2\sigma^2)$

- 如果$X_1, X_2, \cdots, X_n$为X的独立观察结果，且$X \sim N(\mu, \sigma^2)$，则：

  $X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$

练习

让我们为德克解答爱情过山车问题。4个成年人的综合体重小于800磅的概率是多少？假定每个成年人的体重分布都符合N(180, 625)。



练习解答

让我们为德克解答爱情过山车问题。4个成年人的综合体重小于800磅的概率是多少？假定每个成年人的体重分布都符合N(180, 625)。

如果我们用X表示一个成年人的体重，则 $X \sim N(180, 625)$。我们需要先求出4个成年人的体重的分布情况。为了求出这个新分布的均值和方差，我们将X的均值和方差乘以4。于是得出：

$$X_1 + X_2 + X_3 + X_4 \sim N(720, 2500)$$

为了求出 $P(X_1 + X_2 + X_3 + X_4 < 800)$，我们先求标准分：

$$z = \frac{x - \mu}{\sigma} = \frac{800 - 720}{50} = \frac{80}{50} = 1.6$$

在标准正态概率表中查看这个数值，得到0.9452，即：

$$P(X_1 + X_2 + X_3 + X_4 < 800) = 0.9452$$

打断一下……


WHO WANTS TO WIN A
SWIVEL CHAIR
大家好，欢迎再次观看统计邦热门智力节目“转椅赢赢赢”，今晚的节目更精彩。
ANTS TO WIN A



我们今天为你准备了更多搞怪难题，让我们继续加油。在这一轮节目中，我打算问你40个问题，你需要答对30题以上才能进入下一轮比赛，要不就领了鼓励奖离场。每一个问题有四个备选答案。这一轮的标题是“懂我愈发多一些”。祝你好运！

动动笔

以下是第三轮比赛的前5题，都是关于节目主持人的。

1. 他喜欢看哪一部电影？

- **A:** 铁道游击队
- **B:** 少林寺
- **C:** 倩女幽魂
- **D:** 达芬奇密码

2. 他的猫喜欢看哪一部电影？

- **A:** 海底总动员
- **B:** 龟兔赛跑
- **C:** 捕鼠记
- **D:** 惊弓之鸟

3. 他平均每个月花多少钱配衣服？

- **A:** 1,000美元
- **B:** 2,000美元
- **C:** 3,000美元
- **D:** 4,000美元

4. 他多久理发一次？

- **A:** 1个月
- **B:** 2个月
- **C:** 3个月
- **D:** 4个月

5. 他喜欢哪个网站？

- **A:** www.fatdanscasino.com
- **B:** www.gregs-list.net
- **C:** www.you-cube.net
- **D:** www.starbuzzcoffee.com

接着玩，还是转身走？

和以前一样，你不可能这么了解节目主持人，以至于能够答对有关他的所有问题，看来你又要随机回答问题了。

那么，在40个问题中答对30个问题以上的概率是多少呢？我们将根据这个概率决定是去还是留。

动动笔

你该怎样求出在40个问题中至少答对30个问题的概率？要经过哪些步骤才能得出正确答案？如何求均值和方差？

我们并不要求你算出概率，你只要说出求解步骤就行了。

动动笔解答

你该怎样求出在40个问题中至少答对30个问题的概率？要经过哪些步骤才能得出正确答案？如何求均值和方差？

我们并不要求你算出概率，你只要说出求解步骤就行了。

共有40道题目，也就是说共有40次试答机会，每一次试答或是答对，或是答错。而且，我们想求出答对一定数量的题目的概率，为此需要使用二项分布。令$n=40$，由于每个问题都有4个候选答案，所以p为1/4，即0.25。

如果X为我们答对的题数，则我们要求的是$P(X>30)$，即我们必须将$P(X=30)$直至$P(X=40)$的概率算出来，再加总。

我们可以用n,p和q算出均值和方差，其中$q = 1-p$。均值等于np，方差等于npq，于是得出均值 $= 40 \times 0.25 = 10$，方差 $= 40 \times 0.25 \times 0.75 = 7.5$。



使用二项分布会带来繁重的工作。

为了求出答对30题以上的概率，我们需要把11个单独算得的概率加起来——其中的每一个概率都来之不易，计算过程中极易出错。

我们需要找到一个更简便的算法计算二项分布。


要是别的分布也像正态分布一样容易计算，那就美呆了，可我知道这不过是痴人说梦罢了……

正态分布出手相救

我们已经看出，二项分布会让我们的日子不好过，计算繁复艰深且容易出错，时间哗啦啦流逝，换来的却是错误的答案。

似乎绝望了？别担心，还是有容易的办法的。

在某些情况下，可以用正态分布近似代替二项分布。



在某些情况下，泊松分布可以近似代替二项分布，不过，在另一些情况下，正态分布也可以近似代替二项分布。

懂得用其他分布近似代替二项分布十分有用，它能化繁为简。在某些情况下，泊松分布可以帮助我们计算一些繁杂难解的概率。

在另一些情况下，则可以利用正态分布近似代替二项分布。这样做好处极大，我们可以用正态概率表方便地查找需要求解的概率，从而免去种种计算。

我们只需弄清楚在哪些情况下适合进行这种替代就行了。

动动脑

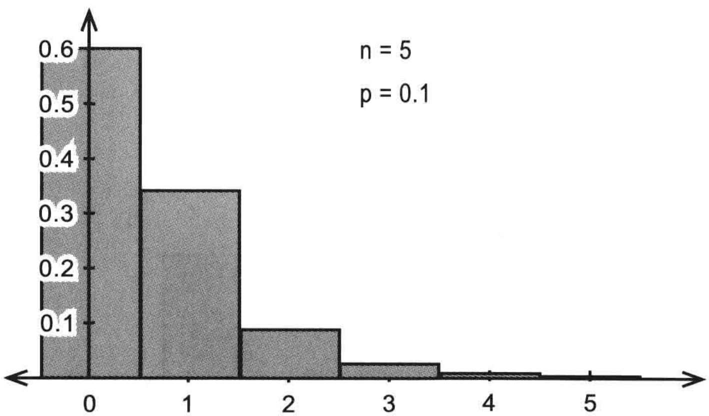
我们在此前一段时间讲过如何使用泊松分布近似代替二项分布，在哪种情况下适合进行这种代替？

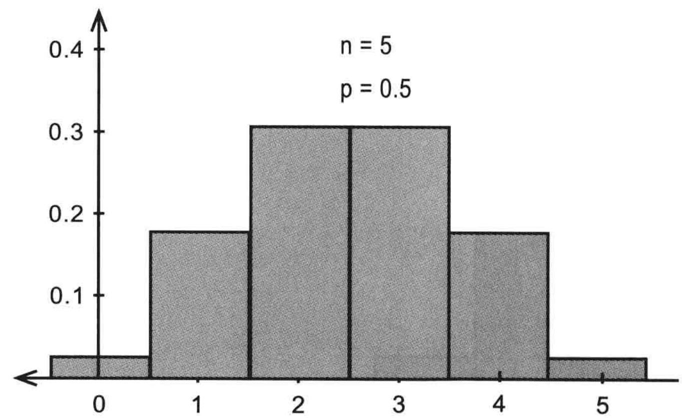
当 $n > 50$ 且 $p < 0.1$ 时，$B(n, p)$ 可以近似用二项分布代替。

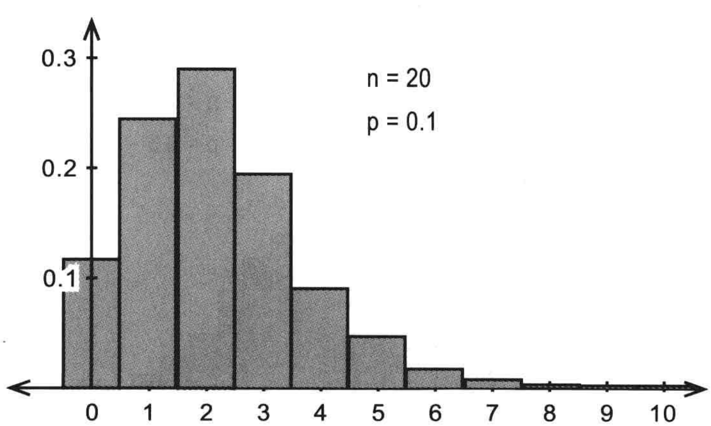


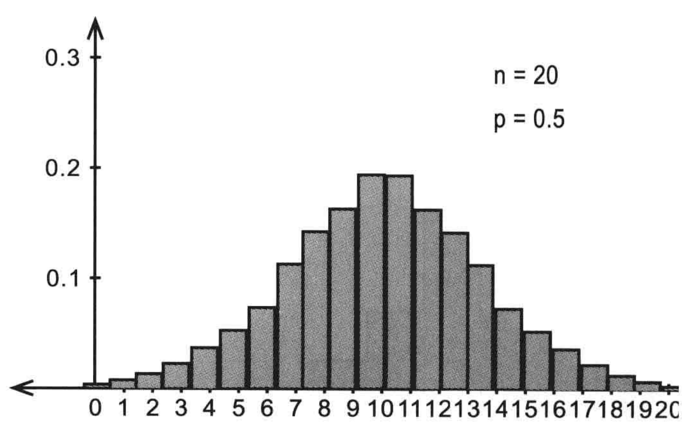
化身分布

以下是一些二项分布，n和p数值各异。你的任务是假装自己是其中的分布，并说出哪一个分布最适合用正态分布进行近似代替。仔细观察每种分布的形状，说说哪一个图形最符合正态。





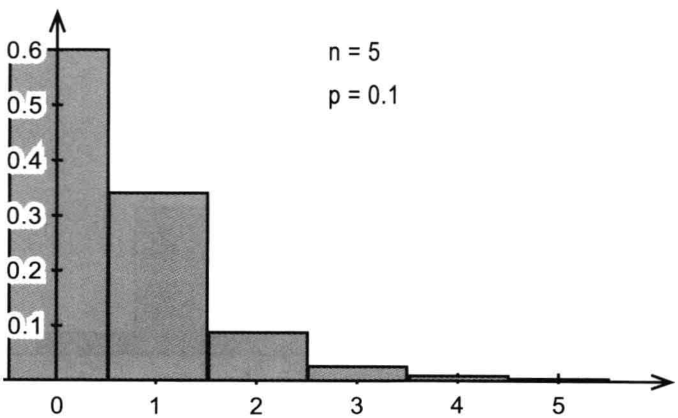




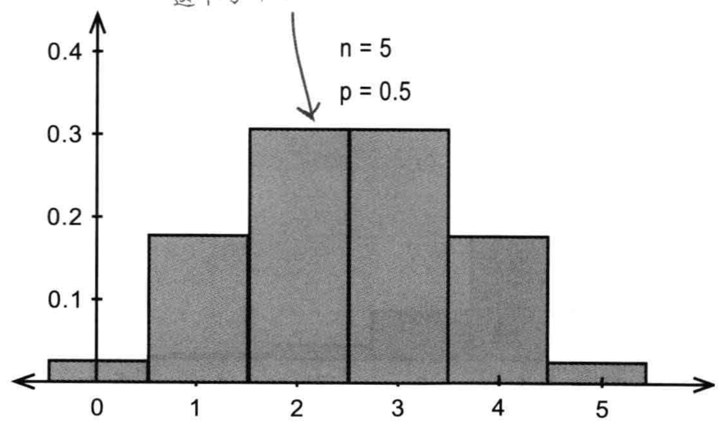
化身分布

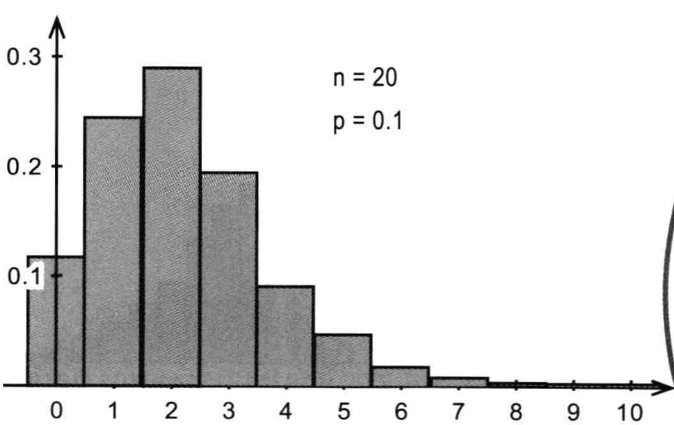


以下是一些二项分布，n和p数值各异。你的任务是假装自己是其中的分布，并说出哪一个分布最适合用正态分布进行近似代替。仔细观察每种分布的形状，说说哪一个图形最符合正态。

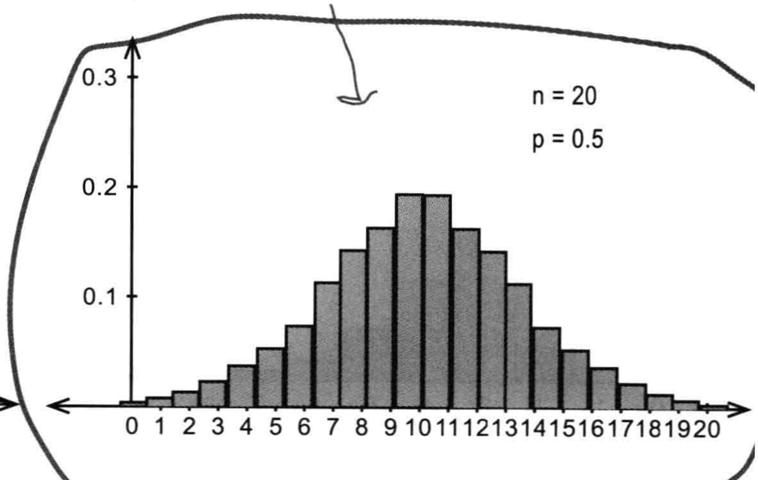


这个分布很对称，但很难取出一段平滑的钟形曲线，用正态分布近似表示这个分布不可能具有足够的正确性。





在这几种分布中，这个分布最适合用正态分布近似代替，当n = 20且p = 0.5时，分布形状与正态分布的形状最为相似。



何时用正态分布近似代替二项分布

在某些情况下，二项分布的形状看上去和正态分布的形状十分相似，在这样的情况下，我们可以用正态分布代替二项分布，得出与二项分布的概率极其近似的结果。我们可以不再大量计算单个概率，而是在标准概率表中查找整个范围的概率。

那么在哪些情况下可以这么做呢？

n是次数，p是成功概率，q等于1-p。

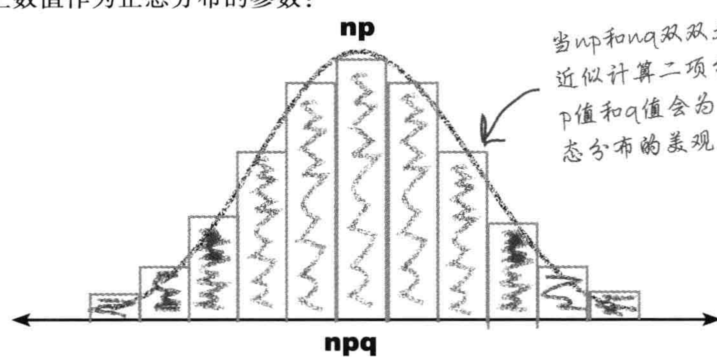
在上一个练习中我们看到，当p在0.5左右、n在20左右时，二项分布的外形与正态分布的外形十分相似，一般说来，当np和nq双双大于5时，可以用正态分布近似代替二项分布。

求解均值和方差

为了能用正态概率表查找概率，我们需要知道均值和方差，以便算出标准差。均值和方差可以直接从二项分布得出，在最初讲到二项分布时，我们发现：

$\mu = np$　　且　　$\sigma^2 = npq$

我们可以把以上数值作为正态分布的参数：



当np和nq双双大于5时，可以用正态分布近似计算二项分布的概率。这样的n值、p值和q值会为我们呈现一个十分接近正态分布的美观、平滑的形状。

重要统计量

二项分布的近似

如果$X \sim B(n, p)$，且$np > 5$，$nq > 5$，则可以使用$X \sim N(np, npq)$近似代替二项分布。



某些课本的近似条件为np > 10及nq > 10。

如果你即将参加统计学考试，一定要问清楚考试委员会的要求。



强化练习

在应用正态分布解决“转椅赢赢赢”的40个问题之前，让我们先用一个简单问题验证一下这种方法的有效性。让我们试着算一算：在12个问题中答对5题或5题以下的概率，其中每个问题只有两个备选答案。

让我们首先用二项分布进行计算，即求出$P(X < 6)$，其中$X \sim B(12, 0.5)$。

现在，让我们用二项分布的正态近似法进行计算，看看是否能得出相同答案。首先，如果$X \sim B(12, 0.5)$，我们可以用哪个正态分布进行近似计算？弄清楚这个问题后，请问$P(X < 6)$是多少？

强化练习解答

在应用正态分布解决“转椅赢赢赢”的40个问题之前，让我们先用一个简单问题验证一下这种方法的有效性。让我们试着算一算：在12个问题中答对5题或5题以下的概率，其中每个问题只有两个备选答案。

让我们首先用二项分布进行计算，即求出$P(X < 6)$，其中$X \sim B(12, 0.5)$。

各个概率用下列公式进行计算：

$$P(X = r) = {}^{n}C_{r}p^{r}q^{n-r} \qquad \text{其中} \qquad {}^{n}C_{r} = \frac{n!}{r!(n-r)!}$$

我们需要求$P(X < 6)$，其中$X \sim B(12, 0.5)$。为此，需要求$P(X = 0)$至$P(X = 5)$，然后将算得的所有概率加起来。

各个概率为：

$$P(X = 0) = {}^{12}C_{0} \times 0.5^{12} = 0.5^{12}$$

$$P(X = 1) = {}^{12}C_{1} \times 0.5 \times 0.5^{11} = 12 \times 0.5^{12}$$

$$P(X = 2) = {}^{12}C_{2} \times 0.5^{2} \times 0.5^{10} = 66 \times 0.5^{12}$$

$$P(X = 3) = {}^{12}C_{3} \times 0.5^{3} \times 0.5^{9} = 220 \times 0.5^{12}$$

$$P(X = 4) = {}^{12}C_{4} \times 0.5^{4} \times 0.5^{8} = 495 \times 0.5^{12}$$

$$P(X = 5) = {}^{12}C_{5} \times 0.5^{5} \times 0.5^{7} = 792 \times 0.5^{12}$$

将以上概率加起来，得到总概率为：

$$P(X < 6) = (1 + 12 + 66 + 220 + 495 + 792) \times 0.5^{12}$$

$$= 1586 \times 0.5^{12}$$

$$= 0.387 \text{（保留三位小数）}$$

现在，让我们用二项分布的正态近似法进行计算，看看是否能得出相同答案。首先，如果X ~ B(12, 0.5)，我们可以用哪个正态分布进行近似计算？弄清楚这个问题后，请问P(X < 6)是多少？

X ~ B(12, 0.5)，即n = 12, p = 0.5，q = 0.5，恰当的近似分布为X ~ N(np, npq)，也就是X ~ N(6, 3)。

我们要求P(X < 6)，所以先计算标准差：

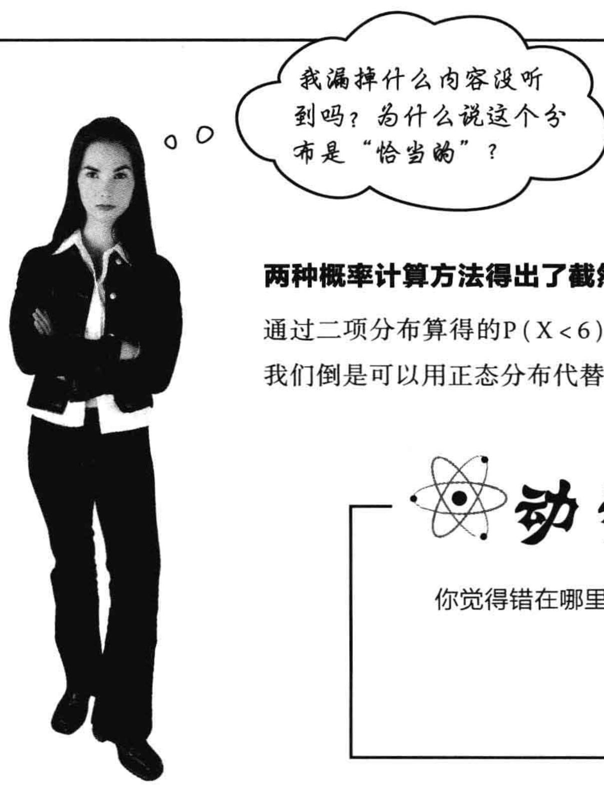
$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{6 - 6}{\sqrt{3}}$$

$$= 0$$

查概率表，得：

P(X < 6) = 0.5



两种概率计算方法得出了截然不同的结果。

通过二项分布算得的P(X<6)等于0.387，而通过正态分布算得的结果为0.5。我们倒是可以用正态分布代替二项分布，但是，结果不够接近。

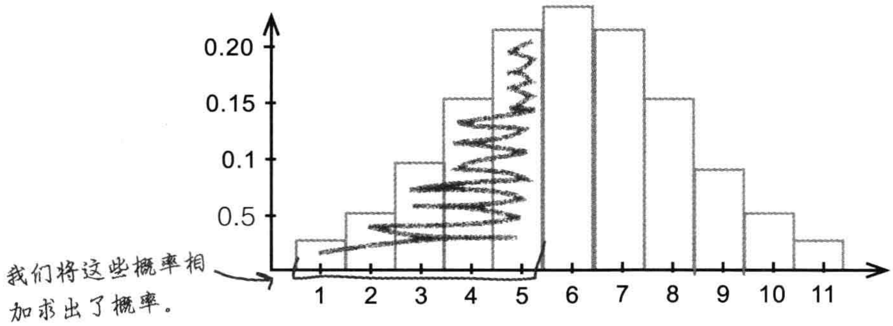
动动脑

你觉得错在哪里呢？我们该如何进行修正？

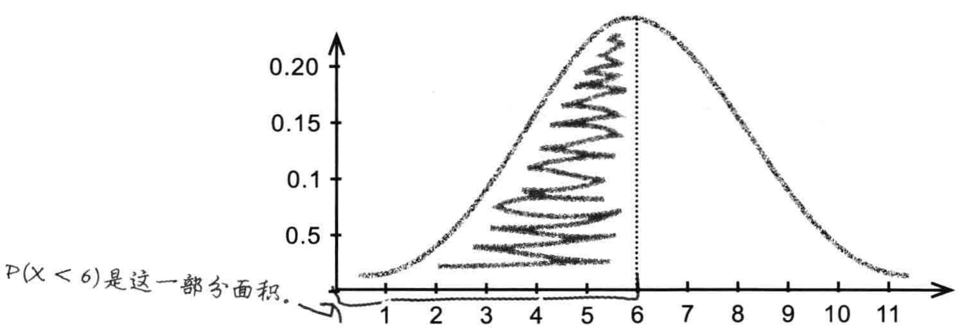
再谈正态近似

错在哪里？让我们仔细研究这个问题，看看能否发现蹊跷，能否想出办法进行处理。

首先看概率分布X ~ B(12, 0.5)，我们想求出答对的问题不足6个的概率，并已通过计算P(X < 6)获得答案。



然后我们用X ~ N(6, 3)对这个分布进行近似，根据需要，为了求出二项分布的概率P(X < 6)，我们用正态分布计算P(X < 6)：



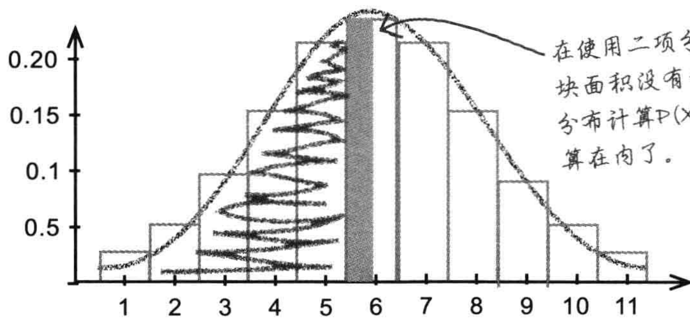
进一步仔细观察两种概率分布。虽然不易察觉，但两者之间确实存在重大差别：我们分别用于计算两个概率的两个范围略有不同。在计算正态分布的时候，我们使用的实际范围略微大一些，这正是概率变大的原因。

下一页将详细讲解这个问题。

二项分布是离散分布，正态分布则是连续分布

我们在对前面的两种概率进行计算时忽略了一件事——没有考虑到其中一种分布是离散分布（二项分布），而另一种分布是连续分布（正态分布）。这很重要，因为我们所用的概率范围会大大影响最终概率。

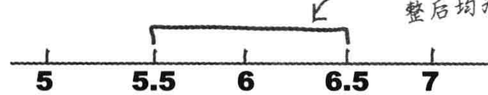
以下在同一张图上体现了X ~ B(12, 0.5)和N(6, 3)这两种概率分布。我们特别指出了正态分布所用概率范围超出二项分布所用范围的部分。



你能看出问题所在吗？

当我们从一个离散概率分布中取出一些整数，并将这些整数转化为连续标度时，我们所观察的并不仅仅是那些精确的孤立数值，相反，我们观察的是由多个数字形成的范围，这些数字经过取整，得到的正是我们取用的那些精确的离散整数。

让我们以离散数值6为例，当我们将数字6转化成一个连续标度时，我们需要考虑所有取整后等于6的数字，即，从5.5到6.5的整个数字范围。



这对于我们的概率问题有什么影响呢？

此前我们试着用正态分布近似计算答对题数在6以下的概率时，没有注意到离散数值6转变成了连续标度。可实际上，离散数值6包含了从5.5到6.5之间的一个范围，因此，我们不应该计算P(X < 6)，而应该试着计算P(X < 5.5)。

这种调整被称为**连续性修正**。在将离散数值转换为连续标度时，所作的小幅调整就是连续性修正。

在计算近似值之前先进行连续性修正

让我们试着求出P(X < 5.5)，其中X ~ N(6, 3)，看看这个概率与答对5题或5题以下的概率的近似程度如何。之前我们已经利用二项分布求出目标概率为0.387左右。

让我们看看正态分布得到的结果的近似程度有多大。

我们想求P(X < 5.5)，其中X ~ (6, 3),让我们先算标准分。

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{5.5 - 6}{\sqrt{3}}$$

$$= -0.29 \text{ (保留两位小数)}$$

看看这两个概率，的确十分近似，看来连续性修正成功了。

我们想求面积Z < -0.29的概率，于是查找标准正态概率表，得到概率为0.3859。即：
P(X < 5.5) = 0.3859

这个概率和我们用二项分布求得的概率十分近似——之前用二项分布算得的概率为0.387，因此正态分布得到的是十分近似的结果。

要 点

- 在一些特定情况下，可以用**正态分布近似代替二项分布**。如果X~B(n, p)，且np > 5，nq > 5，则可以用X~N(np,npq)近似代替X。
- 如果用正态分布近似代替二项分布，则需要**进行连续性修正**，这才能确保得到正确的结果。

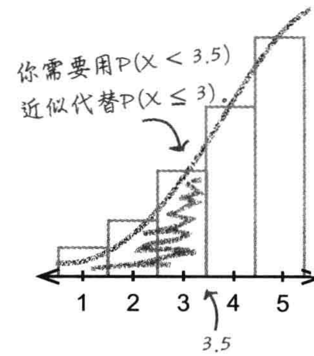
连续性修正细细看

使用正态分布近似代替二项分布有一个技巧，即务必进行合适的连续性修正。如上所见，所选概率范围的小小变化会导致实际得到的概率出现重大误差。听起来这似乎不是什么了不起的大问题，可是，使用错误的概率将会导致做出错误的决策。

让我们看看针对各种概率问题需要使用的各种连续性修正。

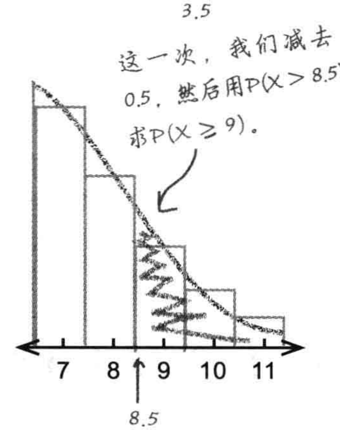
≤型概率的求解

在计算$P(X \le a)$这种形式的概率时，关键是要确保所选择的范围中包含离散数值a。在一个连续标度上，离散数值a会增长到$(a+0.5)$。这就是说，如果使用正态分布求$P(X \le a)$，则实际上需要计算$P(X < a+0.5)$，以此得出近似值，换句话说，你要增加一个额外的0.5。



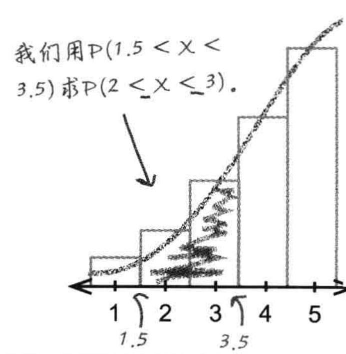
≥型概率的求解

在计算$P(X \ge b)$这种形式的概率时，一定要确保所选择的范围中包含离散数值b。在一个连续标度上，离散数值b会减小到$(b-0.5)$。这就是说，你需要使用范围$P(X > b-0.5)$，这样才能确保该数值位于这个范围内，换句话说，你需要减去一个额外的0.5。



"介于"型概率的求解

在计算$P(a \le X \le b)$这种形式的概率时，需要进行连续性修正，以便确保a和b均包含在内。为此需要将两端的范围均扩展0.5。为了使用正态分布近似计算这个概率，我们需要求$P(a-0.5 < X < b+0.5)$，这正好是以上两种概率类型的综合。



世上没有傻问题

问： 用正态分布近似计算二项分布的确能节省时间吗？

答： 可以节省大量时间。计算二项概率时，通常必须计算大量数值的概率，因此十分费时，没有什么方法能够简便地计算一个数值范围内的所有二项概率。

如果用正态分布近似计算二项分布，那就快多了，你可以在标准表中查找概率，一口气把整个数据范围的概率算出来。

问： 确实能得到精确结果吗？

答： 没错，在大多数情况下都足够精确。但要记住：需要进行连续性修正。如果不进行连续性修正，则结果的正确性将下降。

问： 怎么对<和>进行连续性修正？像≤ 和 ≥一样进行处理吗？

答： 有差别的，这要看你要包含哪个数值，要排除哪个数值。

在用≤ 和 ≥计算概率的时候，你需要确保不等式中的数值落在已知概率范围之内。因此，假如要计算$P(X \leq 10)$，则需要确保数值范围中包含10，即需要考虑$P(X < 10.5)$。

在用<或>计算概率时，你需要确保不等式中的数值落在已知概率范围之外。即，假如要计算$P(X < 10)$，则需要确保数值范围中不包含10，即需要考虑$P(X < 9.5)$。

问： 正态分布和泊松分布都能作为二项分布的近似，我该用哪一个？

答： 这要看具体情况。如果$X \sim B(n, p)$，当$np > 5$且$nq > 5$时，则使用正态分布近似代替二项分布。

如果$n > 50$且$p < 0.1$，则可以使用泊松分布近似代替二项分布。

记住：在用正态分布近似代替二项分布时，必须进行连续性修正。

奇妙池



你的**任务**是从奇妙池中捞出公式因子，将这些因子放入计算式中的横线上，目的是为每一种离散概率范围提供正确的连续性修正。同一因子**可以**多次使用，不必使用所有因子。

$X < 3$ → ____________

$X > 3$ → ____________

$X \leqslant 3$ → ____________

$X \geqslant 3$ → ____________

$3 \leqslant X < 10$ → ____________

$X = 0$ → ____________

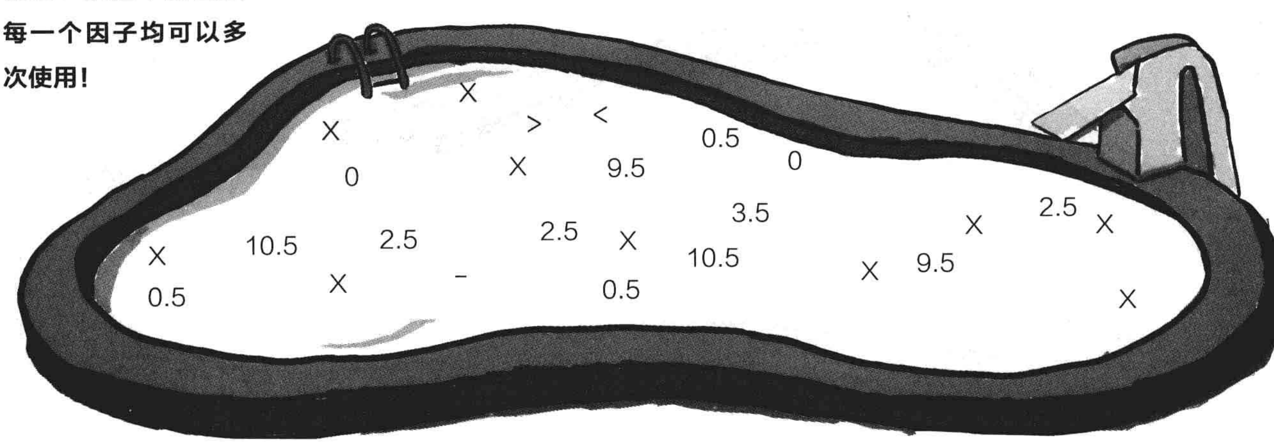
$3 \leqslant X \leqslant 10$ → ____________

$3 < X \leqslant 10$ → ____________

$X > 0$ → ____________

$3 < X < 10$ → ____________

说明：从池中捞出的每一个因子均可以多次使用！



奇妙池解答



你的**任务**是从奇妙池中捞出公式因子，将这些因子放入计算式中的横线上，目的是为每一种离散概率范围提供正确的连续性修正。同一因子**可以**多次使用，不必使用所有因子。

这个式子表示，我们要找出小于3的数值。2.5取整等于3，因此，我们只想让数值范围中包含小于2.5的数。

从-0.5到0.5的所有数取整后都等于0，因此，必须将这些数值包含在数值范围内。

$X < 3 \rightarrow$ $X < 2.5$

$X > 3 \rightarrow$ $X > 3.5$

$X \leqslant 3 \rightarrow$ $X < 3.5$

在这个式子中，我们所求的是小于等于3的数值，2.5到3之间的数值取整后都等于3，因此需要将小于3.5的数值包含在数值范围中。

$X \geqslant 3 \rightarrow$ $X > 2.5$

$3 \leqslant X < 10 \rightarrow$ $2.5 < X < 9.5$

$X = 0 \rightarrow$ $-0.5 < X < 0.5$

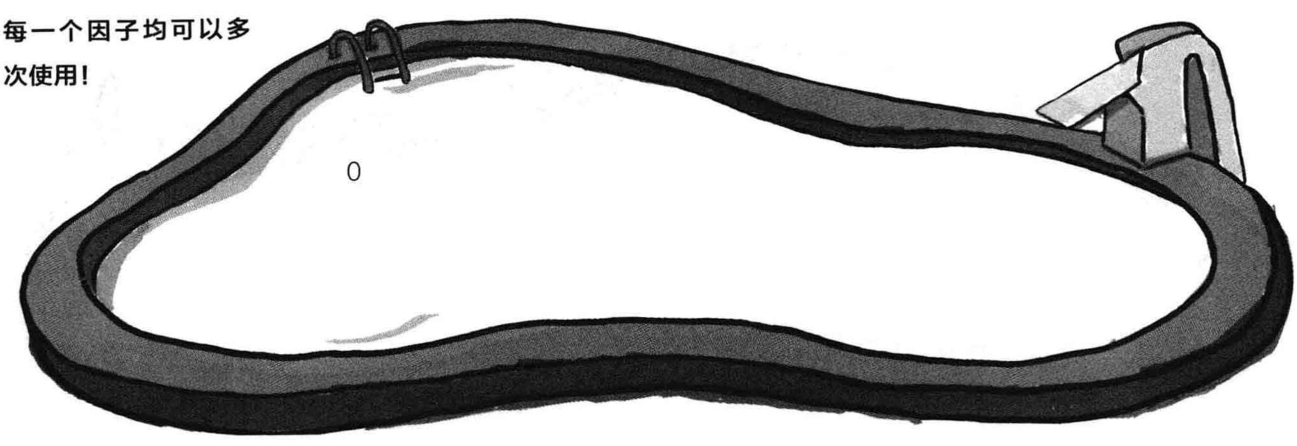
$3 \leqslant X \leqslant 10 \rightarrow$ $2.5 < X < 10.5$

$3 < X \leqslant 10 \rightarrow$ $3.5 < X < 10.5$

$X > 0 \rightarrow$ $X > 0.5$

$3 < X < 10 \rightarrow$ $3.5 < X < 9.5$

说明：从池中捞出的每一个因子均可以多次使用！



练习

在今天这一期“转椅赢赢赢”节目中，你赢得累计奖金的概率有多大？看看你能不能求出在40个问题中答对30题的概率，每个问题有4个备选答案。

练习解答

在今天这一期“转椅赢赢赢”节目中，你赢得累计奖金的概率有多大？看看你能不能求出在40个问题中答对30题的概率，每个问题有4个备选答案。

如果X为答对的问题的数目，那么我们要求$P(X \geq 30)$，其中$X \sim B(40, 0.25)$。

由于np与nq均大于5，所以适合用正态分布近似计算这个概率。$np = 10$，$npq = 30$，于是我们需要求$P(X > 29.5)$，其中$X \sim N(10, 30)$。

让我们先求标准分：

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{29.5 - 10}{30}$$

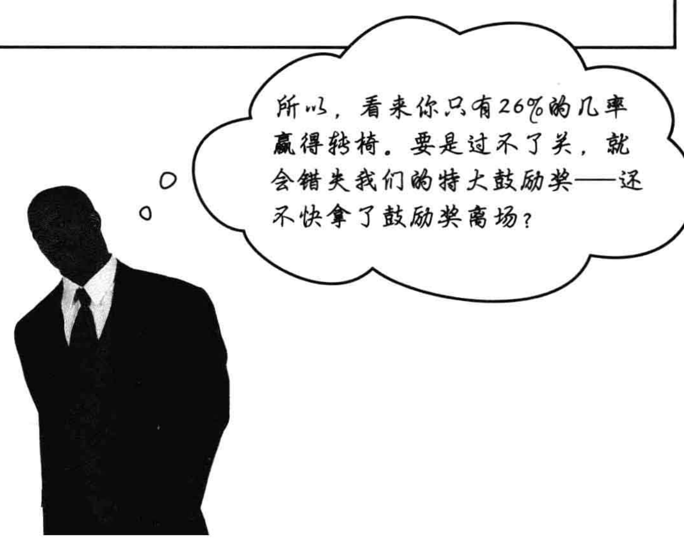
$$= \frac{19.5}{30}$$

$$= 0.65$$

在概率表中查找0.65，得到概率0.7422。即：

$$P(X > 29.5) = 1 - 0.7422$$

$$= 0.2578$$



看你离去让人心伤，你能参加比赛真的很棒。不过，我们刚刚收到一份电子邮件，发件人名叫德克……



动动笔解答

以下是第三轮比赛的前5题，都是关于节目主持人的。

1. 他喜欢看哪一部电影？

- [x] **A:** 铁道游击队
- [ ] **B:** 少林寺
- [ ] **C:** 倩女幽魂
- [ ] **D:** 达芬奇密码

2. 他的猫喜欢看哪一部电影？

- [ ] **A:** 海底总动员
- [ ] **B:** 龟兔赛跑
- [x] **C:** 捕鼠记
- [ ] **D:** 惊弓之鸟

3. 他平均每个月花多少钱配衣服？

- [ ] **A:** 1,000美元
- [ ] **B:** 2,000美元
- [ ] **C:** 3,000美元
- [x] **D:** 4,000美元

4. 他多久理发一次？

- [ ] **A:** 1个月
- [x] **B:** 2个月
- [ ] **C:** 3个月
- [ ] **D:** 4个月

5. 他喜欢哪个网站？

- [x] **A:** www.fatdanscasino.com
- [ ] **B:** www.gregs-list.net
- [ ] **C:** www.you-cube.net
- [ ] **D:** www.starbuzzcoffee.com



组合访谈

本周话题：
为什么“正”不等于“闷”

Head First：嗨，正态兄，真高兴你能来参加节目。

Normal：谢谢你邀请我，Head First。

Head First：现在，我的第一个问题与你的名字有关。你为什么叫做“正态”？

Normal：“正态”是中文说法，其实，在英语里，我的名字是“normal”，意思是“常见的，典型的”，主要是因为我能恰当代表多种多样的数据类型。这些数据的概率分布具有独特的形状——钟形，十分平滑，这正是我。我可以说是理想型吧。

Head First：可以举一个例子吗？

Normal：当然可以，假设你开了一家点心店，店里出售各种面包。理论上每一块特定品种的面包都应该重量相同，但实际上每一块面包的确切重量会有波动。

Head First：不过，这些面包称起来肯定一样重吧？

Normal：大致一样，但存在偏差。我为这种偏差建模。

Head First：建立模型为什么这么重要？

Normal：嗯，这表示你可以用我来计算概率。假设你随机选取一块面包，要计算这块面包的重量小于某个特定值的概率——这听起来像是十分难办，不过，有我在就简单了。

Head First：简单？你指的是？

Normal：其他许多概率分布会牵涉到大量错综复杂的计算。二项分布需要使用阶乘；泊松分布需要计算幂指数，而我不用算这些。只要在概率表中查一查，就解决了。

Head First：肯定没这么容易吧？

Normal：哦，首先要把我转化成标准分，不过这不足挂齿，无碍大局。

Head First：告诉我，你是否觉得自己比别的概率分布都强一些？

Normal：我不会说我比别的概率强多少，不过我倒是灵活许多，在很多地方都派得上用场。我也更健全，当泊松分布和二项分布的数字变得很大时，他们就会遇上麻烦。话说回来，我会尽力帮忙的。

Head First：是吗？怎么帮呢？

Normal：哦，在某些情况下，二项分布和泊松分布看上去都和我相似，这一点颇为诡异。在聚会上，常常有人拦住他们，问他们是不是正态先生，我对他们说，就当别人在恭维你们吧。

Head First：这能带来什么帮助呢？

Normal：哦，由于他们看上去像我，实际上就可以用我的概率表算出他们的概率。用处有多大？那就是再也不用深更半夜地拿计算器了，只需一个字：查。

Head First：由于时间关系，看来今天只能谈到这儿了。正态先生，谢谢你的到来，采访你真愉快。

Normal：别客气，Head First。

大家坐上爱情过山车

还记得德克的爱情过山车吗？他已经开始请人试坐，每一个试坐过的人都觉得很棒。只有一个问题：过山车有时候会发生故障，故障导致延迟，延迟导致耗钱。

关于正在试用的这款过山车，德克在网上找到了一些统计数据，其中一个网站说可以预期的故障次数为每年40次。



看在过山车肯定能赚大钱的份上，德克考虑，如果过山车的停机概率低于每年52次，还是值得干下去的。

我们如何算出这个概率呢？

动动笔

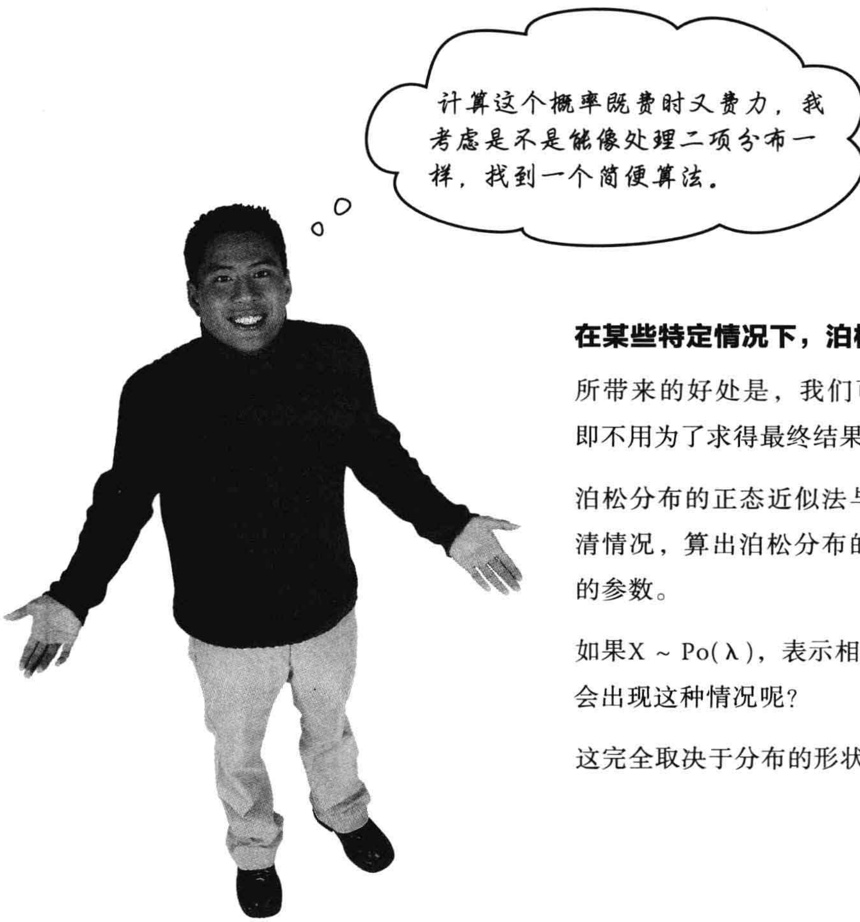
这种情况符合哪种概率分布？如何求出过山车每年发生的故障小于52次的概率？

动动笔解答

这种情况符合哪种概率分布？如何求出过山车每年发生的故障小于52次的概率？

如果某物体以某种平均频率发生故障，则这种情况符合泊松分布，以均值为其参数。如果X表示一年内的故障次数，则X ~ Po(40)。

我们需要求P(X < 52)，为此，我们需要求出52以内的所有X值分别对应的概率。



在某些特定情况下，泊松分布的形状很像正态分布。

所带来的好处是，我们可以利用标准正态概率表算出全部概率，即不用为了求得最终结果而大量计算一个个概率。

泊松分布的正态近似法与二项分布的正态近似法十分相似：先认清情况，算出泊松分布的均值和方差，然后将二者作为正态分布的参数。

如果X ~ Po(λ)，表示相应的正态近似为X ~ N(λ，λ)。什么时候会出现这种情况呢？

这完全取决于分布的形状。

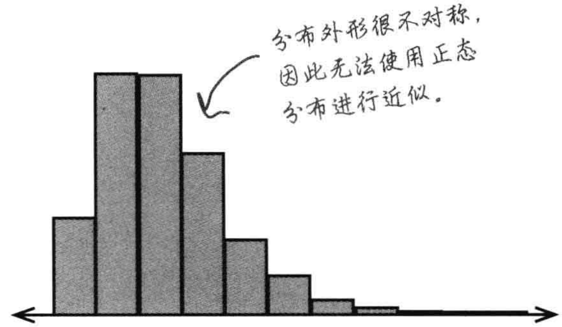
何时能用正态分布近似代替泊松分布

只要泊松分布的形状与正态分布相似，就可以用正态分布近似代替泊松分布。什么时候会出现这种情况呢？让我们看看。

当λ很小……

当λ很小时，泊松分布的形状与正态分布不相同，图像不对称，曲线好像被“扯”向了右边。

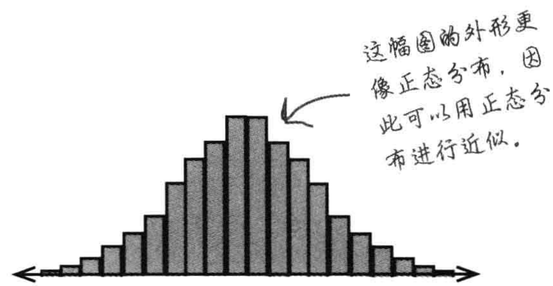
由于泊松分布在λ较小时与正态分布差别很大，因此在λ较小时，不适合用正态分布近似代替泊松分布。



当λ很大……

随着λ变大，泊松分布图的外形看起来越来越像正态分布。曲线的主要部分呈合理对称，近似光滑曲线，与正态分布接近。

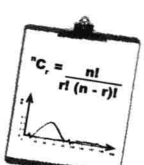
即，随着λ变大，正态分布越来越适合用来近似泊松分布。



多大才算足够大？

我们已经看到，当λ较大时，泊松分布与正态分布相似，不过，要大到什么程度才能用正态分布进行近似呢？

当λ大于15时可谓足够大。即，如果$X \sim Po(\lambda)$且$\lambda > 15$，我们就能用$X \sim N(\lambda, \lambda)$近似计算$X \sim Po(\lambda)$。



重要统计量

泊松分布的近似

如果$X \sim Po(\lambda)$且$\lambda > 15$，则可用$X \sim N(\lambda, \lambda)$进行近似。

练习

德克的爱情过山车发生故障的次数符合泊松分布，其中 $\lambda = 40$。

第一年的故障次数小于52次的概率有多大?

提示：用正态近似法，
别忘了连续性修正。



练习

现在该考考你的统计知识了。填写下表，说说哪种正态分布适合哪种情况，需要满足什么条件。

情况	分布	条件
$X + Y$ **$X \sim N(\mu_x, \sigma^2_x)$, $Y \sim (\mu_y, \sigma^2_y)$**	$X + Y \sim N(\mu_x + \mu_y, \sigma^2_x + \sigma^2_y)$	X与Y为独立变量
$X - Y$ **$X \sim N(\mu_x, \sigma^2_x)$, $Y \sim (\mu_y, \sigma^2_y)$**		
$aX + b$ **$X \sim N(\mu, \sigma^2)$**		
$X_1 + X_2 + \cdots + X_n$ **$X \sim N(\mu, \sigma^2)$**		
X的正态近似 **$X \sim B(n, p)$**		
X的正态近似 **$X \sim Po(\lambda)$**		



德克的爱情过山车发生故障的次数符合泊松分布，其中 $\lambda = 40$。

第一年的故障次数小于52次的概率有多大？

如果用X表示一年内的故障次数，则 $X \sim Po(40)$。

由于 λ 较大，我们可以用正态分布近似代替这个分布，即可以用：

$X \sim N(40, 40)$

我们需要求故障次数小于52的概率。由于用连续概率分布近似代替离散概率分布，所以必须进行连续化修正。我们不应将52计算在内，于是只需要求 $P(X \leq 51.5)$。

在用标准正态表查出概率之前，需要先计算标准分。

$$z = \frac{x - \mu}{\sigma}$$

$$= \frac{51.5 - 40}{6.32}$$

$$= 1.82 \text{（保留两位小数）}$$

在概率表中查找以上结果，得到0.9656，即一年内的故障次数小于52的概率为0.9656。



练习解答

现在该考考你的统计知识了。填写下表，说说哪种正态分布适合哪种情况，需要满足什么条件。

情况	分布	条件
X + Y **$X \sim N(\mu_x, \sigma^2_x)$, $Y \sim (\mu_y, \sigma^2_y)$**	$X + Y \sim N(\mu_x + \mu_y, \sigma^2_x + \sigma^2_y)$	X与Y为独立变量
X - Y **$X \sim N(\mu_x, \sigma^2_x)$, $Y \sim (\mu_y, \sigma^2_y)$**	$X - Y \sim N(\mu_x - \mu_y, \sigma^2_x + \sigma^2_y)$	X与Y为独立变量
aX + b **$X \sim N(\mu, \sigma^2)$**	$aX + b \sim N(a\mu + b, a^2\sigma^2)$	a, b为常量
$X_1 + X_2 + \cdots + X_n$ **$X \sim N(\mu, \sigma^2)$**	$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2)$	$X_1, X_2, \cdots, X_n$ 为X的独立观察结果
X的正态近似 **$X \sim B(n, p)$**	$X \sim N(np, npq)$	$np > 5$, $npq > 5$ 需要进行连续性修正
X的正态近似 **$X \sim Po(\lambda)$**	$X \sim N(\lambda, \lambda)$	$\lambda > 15$ 需要进行连续性修正

要点

- 在特定条件下，可以使用正态分布近似泊松分布。
- 如果$X \sim Po(\lambda)$且$\lambda > 15$，则可以用$X \sim N(\lambda, \lambda)$近似X。
- 如果你用正态分布近似代替泊松分布，那么，为了确保结果正确，需要进行连续性修正。

世上没有傻问题

问： 二项分布和泊松分布都可以用正态分布近似表示，那么几何分布可以吗？正态分布能近似代替几何分布吗？

答： 我们之所以可以用正态分布近似代替二项分布和泊松分布，是因为在某些特定情况下，这两种分布与正态分布具有相同的形状。

而几何分布呢，它永远也不会和正态分布外形相似，因此，正态分布绝不能有效地近似代替几何分布。

问： 如果用正态分布近似代替泊松分布，必须进行连续性修正吗？

答： 没错，这是因为你在用连续概率分布近似代替离散概率分布，因此就像修正二项分布一样，需要对泊松分布进行连续性修正。

问： 用正态分布近似代替二项分布或泊松分布有什么好处呢？如果坚持用原来的分布，结果是不是会更准确呢？

答： 如果使用原来的分布，结果的确会更准确，但这极费时间。如果你想通过二项分布或泊松分布求出一个数值范围的概率，就需要求出该数值范围中的每一个单独数值的概率。相反，使用正态分布则可以查找整个范围的概率，这样就大大地简化了。

用正态分布近似代替泊松分布时，要进行连续性修正。

婚礼成功！

经过你高明的统计分析，爱情过山车开张了，客户需求比德克的最高预期还要旺盛。下面就是德克的一部分顾客，看，他们多幸福！



10 统计抽样的运用

抽取样本



统计需要处理数据，数据从何而来？

有时候数据很容易收集——例如参加一家健身俱乐部的人员的年龄，或一家游戏公司的销售数据。但有时候不太容易，这时候该怎么办？——当事件数量十分庞大时，很难决定该从何处着手收集数据。在本章中，我们将看看如何在实际工作中**成功收集数据**——有效地、正确地、省时省钱地收集数据。欢迎来到抽样天地。

曼帝糖果公司口味检验

曼帝糖果公司是一家糖果和巧克力主要供应商，超长效口香糖球是他们的标志性产品，这种产品具有五彩缤纷的颜色，可以满足各种口味。

曼帝糖果公司打算大做电视广告，吸引更多的消费者，广告包括这样一部分内容：宣传口香糖球的口味持续时间。问题来了：他们该怎样得到相关数据？

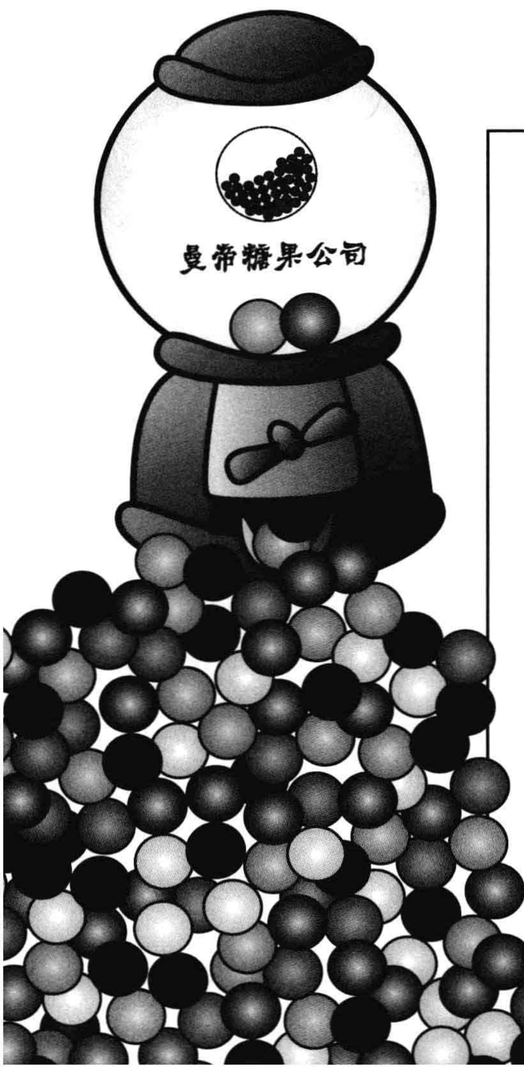
公司决定进行口味检验，也已经招聘了一批试吃者帮忙完成检验。这时出现了两个问题：试吃者吃完了所有的糖球；试吃者的牙齿健康问题让公司破费不少。



糖球吃光了

曼帝糖果公司口味检验发生了重大失误 —— 试吃者把**所有的**糖球都吃光了。这不仅伤害了试吃者的牙齿，而且没有糖球可卖了 —— 试吃者嚼过的糖球是不能拿来卖的。

进行口味检验的目的是弄清楚糖球的口味持续时间，但这真的意味着试吃者必须尝遍每一粒糖球吗？



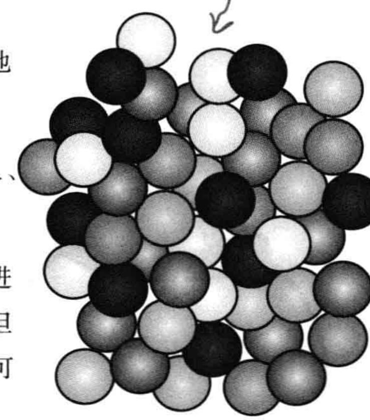
动动脑

为了确定糖球的口味持续时间，你会怎么做？需要考虑什么？将答案写在下面，尽量写详细些。

对糖球样本而非糖球总体进行检验

曼帝糖果之所以碰到问题，是因为他们的试吃检验出现了“试吃每一粒糖球”这个环节，这个环节费时、费钱、伤牙齿，并且剩不下糖球卖给消费者。

那么，曼帝糖果该做些什么改变呢？让我们从总体和样本的差别讲起。



糖球总体

目前，曼帝糖果对现有的每一粒糖球进行口味检验，若用统计术语表达，那么他们是在用**总体**进行检验。

统计学上的**总体**指的是准备对其进行测量、研究或分析的整个群体，可以是人、得分，也可以是糖果 —— 关键在于总体指的是所有对象。

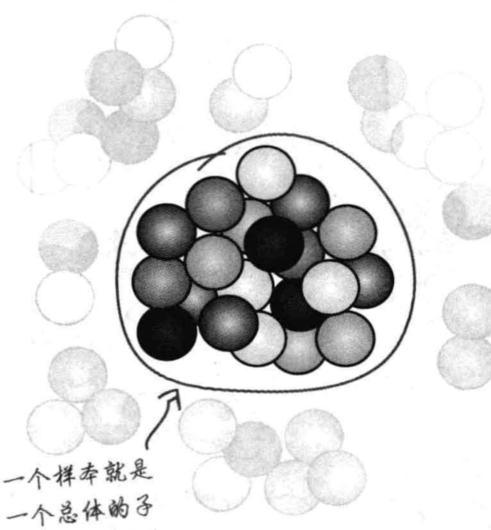
普查指的是对总体进行研究或调查。在曼帝糖果的实例中，他们对每一粒糖球进行品尝，因此，是对糖球总体进行普查。普查可以给出关于总体的准确信息，但并不是在任何情况下都切实可行。当总体数量很大，或者说无穷无尽时，就不可能对每一个对象进行研究了。

糖球样本

不需要尝遍所有糖球也能搞清楚糖球口味持续时间 —— 你可以不检验总体，而检验**样本**。

一个统计**样本**就是从总体中选取的一部分对象。通过选取样本，使其恰当地代表总体，从而得到代表总体的一个子集。对于曼帝糖果来说，一个口香糖球样本就是所选取的一**小部分**糖球，而不是每一粒糖球。

仅对总体的一个样本进行的研究或调查称为**样本调查**，在多数情况下，进行样本调查比进行普查更切实可行，通常样本调查所费的时间和费用都较低，且不用考虑整个总体。由于不使用总体，对口香糖球进行样本调查则意味着调查完毕后还会剩下大量糖球。



那么如何利用样本得出关于总体的结论呢？让我们看一看。

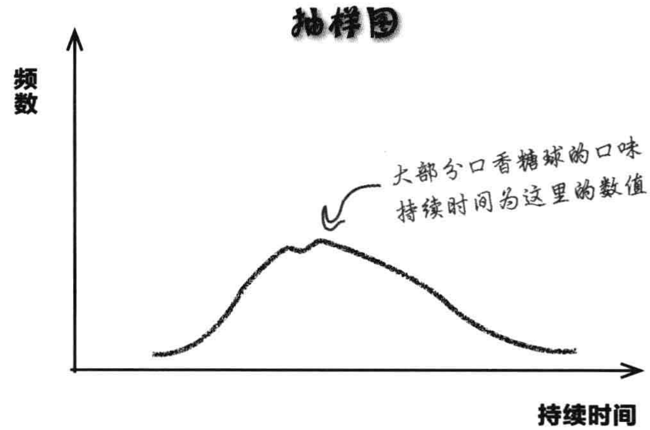
抽样方法

建立一个好样本的关键是尽量选择最符合总体的样本，如果样本具有代表性，则表示样本具有与总体十分相似的特性，进而意味着可以通过样本预测出总体具有哪种特性。

假定你用一个具有代表性的口香糖球样本检验每种口香糖球的口味持续时间，检验结果的分布可能如下所示：

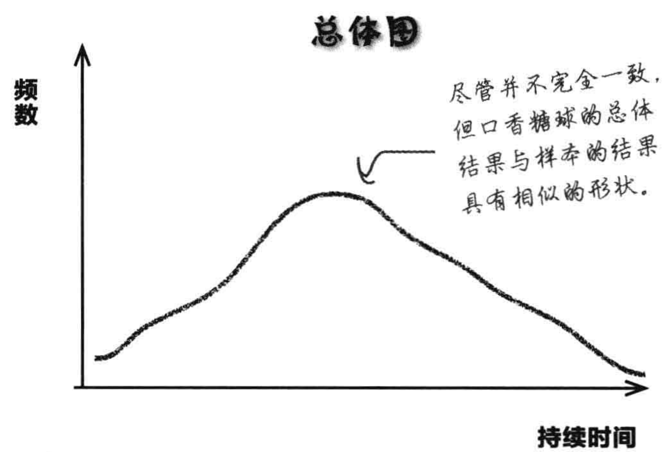
即使只是试吃了一个小样本的口香糖球，你也能对分布形状得出印象。试吃数量越多，图形形状越清晰。例如，通过查看抽样分布的形状，可以对总体分布的中心位置得出初步印象。

让我们将这张图与实际总体进行比较：



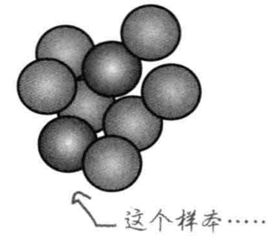
这是总体分布图。看出总体分布和抽样分布有多么接近了吗？

比较这两个图形可以看出，尽管一个图形代表所有的口香糖球，另一个图形仅代表其中一些糖球，但二者的大致形状十分相似。它们具有一些共同的特点——例如数据中心的位置相同，这意味着可以用样本数据预测总体数据。





当抽样有误时

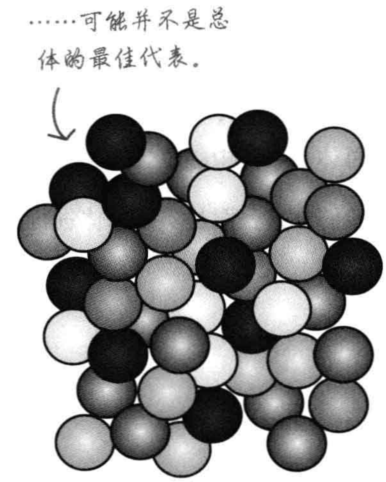


但愿我们能保证每一个样本都与作为样本来源的总体相吻合 —— 可惜，并非每一个样本都酷似其总体。这似乎不是什么大问题，但是，使用具有误导性的样本实际上会导致对总体做出错误的结论。

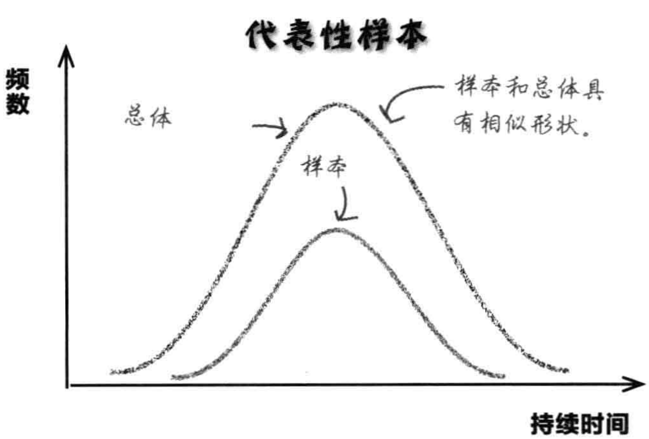
例如，设想你为了检验糖球口味典型持续时间而抽取一个口香糖球样本，但这个样本却仅包含红色糖球，这时，样本可能能够代表红球，却不能代表总体中各种其他颜色的糖球。如果用这个样本的结果推测有关口香糖球总体的信息，最终会对口香糖球的特性形成错误结论。

使用错误的样本会导致对总体参数(例如均值和标准差)得出错误的结论，你可能会对数据形成截然不同的观点，进而做出错误决策。

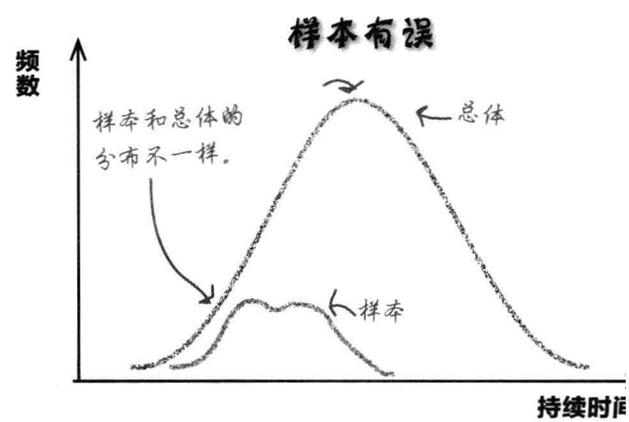
麻烦在于，你可能当局者迷 —— 你可能会觉得总体会如此这般，而事实却并非如此。我们务必建立某种机制，确保样本能够可靠地代表总体。



我们想得到这个结果：



而不是这个结果：



案件：消失的咖啡销量

星巴仕咖啡店首席执行官想在店里销售一种新品牌的咖啡，但他不确定这种咖啡是否受客户欢迎。他让新来的实习生进行调查，摸清客户的想法。实习生请客户品尝新品牌的咖啡，然后把客户的想法告诉首席执行官。



这位实习生十分乐意得到这个大好的工作机会，首先，他已经打听到，如果这个工作干得好，月底将得到一份奖金；其次，他打算向星巴仕的友好客户分发免费咖啡，并聆听一些积极信息；第三，他一直在找借口，想和他常驻的咖啡店的一位常客——一位很特别的女孩搭话，这次工作正是一个机会。

这位实习生做完调查后，兴冲冲地跑去告诉首席执行官人人都喜欢新品牌咖啡，这种新品牌很可能销量火爆。“太好了”，首席执行官说，“我们下个季度就推出这种咖啡。”

当新品牌咖啡最终上市后，销量很不好，首席执行官不得不取消这个系列。你觉得问题出在哪里？

新品牌咖啡为什么销路不佳?

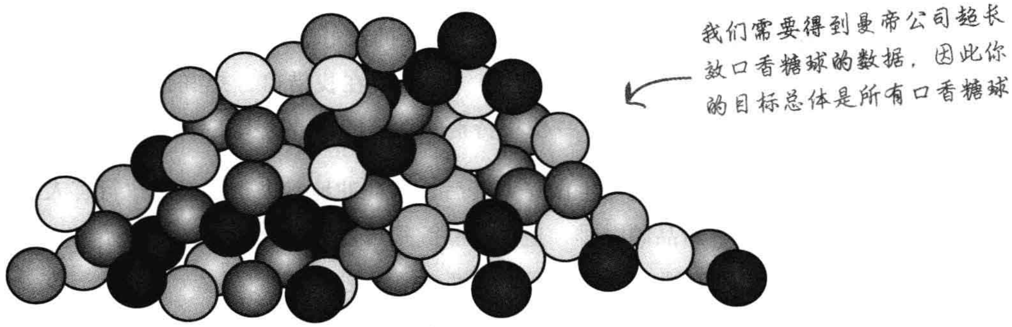
如何设计样本

样本的作用是用它判定总体情况。为了确保得到正确结果，需要明智地选择样本。让我们先来认清总体的实质，以便让样本尽量具有代表性。

确定目标总体

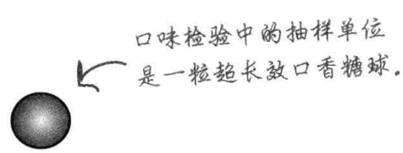
首先要弄清楚**目标总体**何在，才知道样本取自哪里。这里的**目标总体**指的是你正在研究的、并且打算为其采集结果的群体。你所选择的目标总体在很大程度上取决于你的研究目的，例如，你打算收集世界上所有的口香糖球的数据，还是收集某个特定品牌或某个特定类型的口香糖球的数据？

目标总体要尽可能精确，这样能更为容易地得出尽可能代表总体的样本。



确定抽样单位

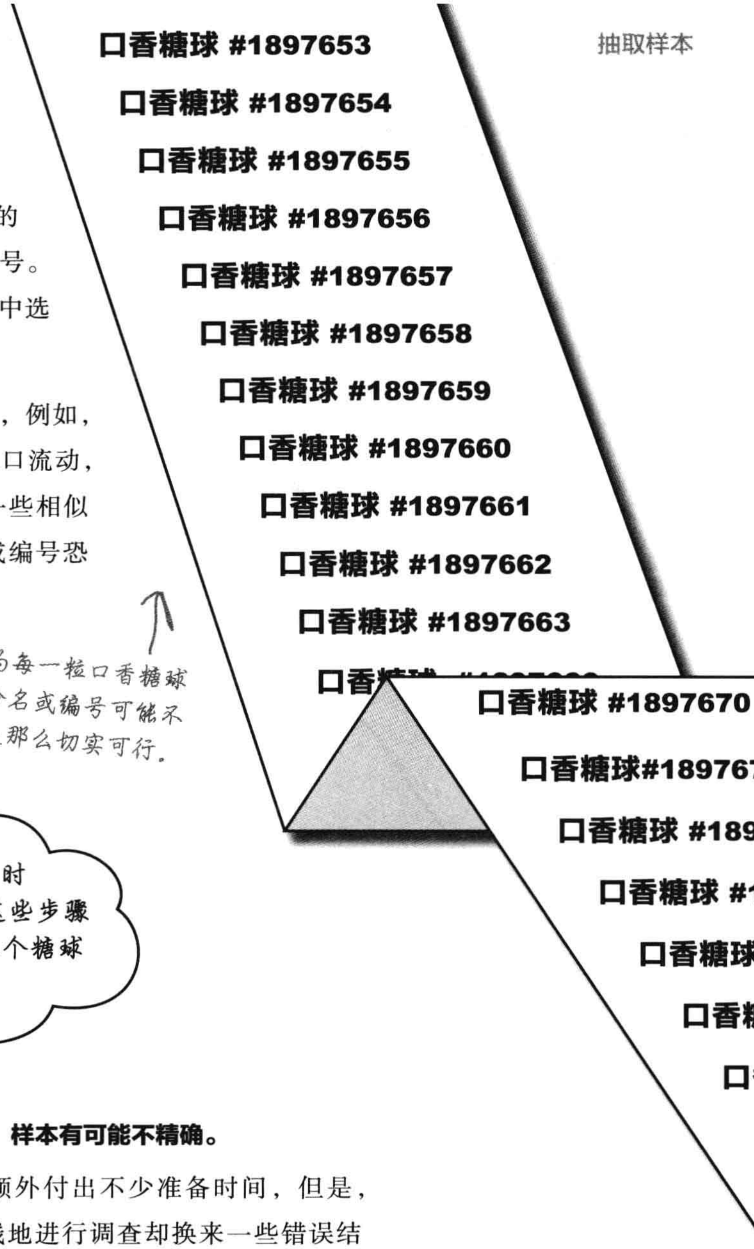
一旦确定目标总体，就需要决定要抽取哪一类对象，通常，要抽样的对象类型就是在确定目标总体时所描述的对象类型，例如，可以是一粒口香糖球，也可以是一盒口香糖球。



确定抽样空间

最后，你需要列一张表，表中列出目标总体范围内的所有抽样单位，最好给每个抽样单位取个名或编个号。这张表被称为**抽样空间**，基本上，你可以从这张表中选取样本。

有时候不可能得出涵盖整个目标总体的抽样空间表，例如，如果要收集生活在某个地区的居民的观点，由于人口流动，表中列举的名字就会受到影响；如果所处理的是一些相似的对象，例如口香糖球，那么为每一粒糖球命名或编号恐怕是不可能的，或者说是不现实的。



为每一粒口香糖球命名或编号可能不是那么切实可行。

这似乎是在浪费时间，我必须完成这些步骤吗？我不能抽取几个糖球样本就完事吗？



如果不好好设计，样本有可能不精确。

设计样本需要额外付出不少准备时间，但是，比起费时、费钱地进行调查却换来一些错误结果，这要好多了。后者会让金钱和时间付诸东流，更有甚者，恐怕会有人根据错误的调查结果做出错误的决策。

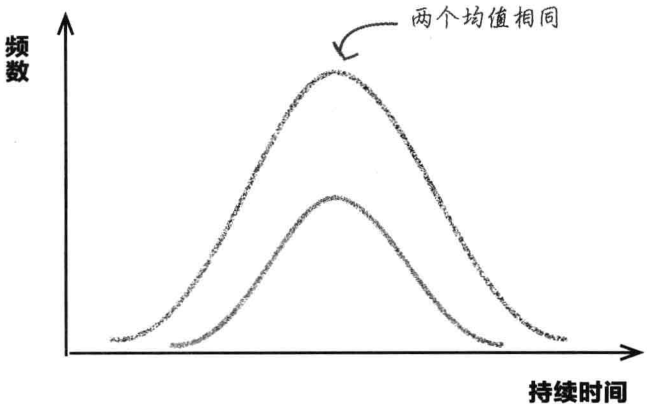
设计不当的样本会引起**偏倚**，让我们详细讲讲这一点。

样本有时会发生偏倚

并非每一个样本都能做到十分客观——除非极其小心，否则，样本中会潜入这样那样的**偏倚**，使最终结果发生扭曲。你在无意间（也可能是有意间）带入样本的某种个人偏好就是**偏倚**，这时，你的样本不再是从总体中进行随机选择的结果。

如果一个样本**无偏**，则这个样本可以代表总体，是总体的客观反映。

无偏样本

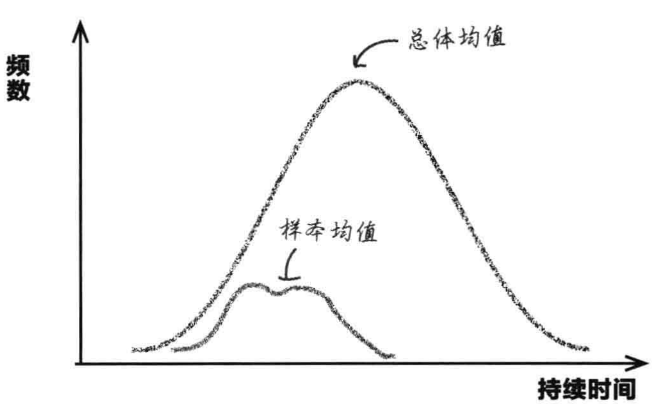


无偏样本

无偏样本可以代表目标总体，即该样本与总体样本具有相似特性，我们可以利用这些相似特性对总体本身做出判断。

一个无偏样本的分布形状与作为其来源的总体的分布形状相似，如果我们知道样本的分布形状，就可以据此以合理程度的置信水平预测总体的分布形状。

偏倚样本



偏倚样本

偏倚样本无法代表目标总体，由于样本与总体的特性不相似，无法根据样本对总体做出判断。如果我们试图用样本的分布形状预测总体的分布形状，最终将会得出错误的结果。



偏倚的来源

偏倚是怎么溜进样本里的？下面是部分原因：

1. 抽样空间中条目不齐全，因此未包含目标总体中的所有对象。如果条目不出现在抽样空间中，那么也不会出现在样本中。

2. 抽样单位不正确。例如，也许抽样单位不应该是一粒粒的口香糖球，而应该是一盒盒的口香糖球。

3. 为样本选取的一个个抽样单位未出现在实际样本中。例如，你可能发出一份调查问卷，但并不是人人都给出回应。

4. 调查问卷的问题设计不当。设计的问题要中性，要适合每个人回答。例如，“曼帝糖果公司的糖果比其他品牌的糖果更可口，您同意吗？”这种提问带有偏倚，较好的做法是请受调查者自己说出他们偏爱的糖果品牌。

5. 样本缺乏随机性。例如，如果在大街上展开调查，你可能会回避行色匆匆或气势汹汹的人，于是你就将气势汹汹的人或行色匆匆的人排除在调查范围以外。



如上所述，偏倚来源广泛，而其中大部分归咎于样本选取方法。

我们需要检查样本的选取方法，使偏倚的发生几率降至最低程度。

世上没有傻问题

问： 这么说抽样空间就是我们所抽取的所有对象的列表？

答： 抽样空间列出总体中的所有独立单位，被作为样本的基础，但它并不是样本本身，这是因为我们不会抽取抽样空间中的所有对象。

问： 我如何形成抽样空间？

答： 具体做法以及所用对象取决于你的目标总体，例如，如果你的目标总体是所有汽车车主，那么可以采用汽车车主花名册；如果你的目标总体是入读某所大学的全体学生，那么可以采用大学注册表。

问： 电话簿之类的东西怎么样？能作为抽样空间吗？

答： 这完全取决于你的目标总体。电话簿上不列出未装电话的家庭，还有一些家庭尽管装了电话，但会选择不在电话簿上公开。如果你的目标总体是有公开的电话号码的家庭，那么使用电话簿是一个不错的主意；如果你的目标总体是所有装有电话的家庭或甚至是所有家庭，那么你的抽样空间不会十分精准——这会带来偏倚。

问： 我总是能拟定抽样空间吗？

答： 并非如此。想象一下，假如你不得不调查海洋中的所有鱼类——为每一条鱼命名、编号是不可能办到的。

问： 我必须确定一个目标总体吗？

答： 不错。你需要知道你的目标总体是什么，这样才能确保样本代表总体。仔细考虑目标总体有助于避免偏倚。

如果你正在替别人做抽样，要尽量搞清楚目标总体是什么。要确保自己确切地知道哪些包含在总体内，哪些排除在总体外。

问： 偏倚为什么如此有害？

答： 偏倚的害处在于会导致对目标总体做出错误结论，进而导致做出错误决策。例如，如果你仅仅抽取粉色口香糖球，对于全部粉色糖球来说，你的调查结果可能是准确的，但对于糖球整体来说却未必准确——不同颜色的糖球之间可能存在重大差异。

问： 调查问卷中的提问如何导致偏倚？

答： 偏倚常常在问题设计阶段悄悄潜入。

首先，如果你给出一系列描述，然后要求受调查者表示同意或不同意，除非受调查者非常反感，否则表示同意的可能性更大。也就是说，调查结果将会偏向同意。

若你给出一组可能答案，但并未涵盖一切可能结果，那么也会出现偏倚。例如，假设你需要向别人提问他们一般一星期锻炼几次，如果你给出的答案是“每星期大于5次”、“每星期3–5次”、“每星期1–2次”以及“我不重视健康，因此不锻炼”，那么就会导致偏倚，因为有些人可能不锻炼身体，但他们并不同意“不重视健康”这一说法，也就是说，他们无法回答问题。

动动笔

考虑下面这些情况，你会选择什么作为目标总体？抽样单位是什么？你会如何拟定抽样空间？进行抽样时还需要考虑哪些问题？

1. 巧口华公司生产巧克力，他们为节庆季度限量生产了一些巧克力，想要检验这些巧克力的品质。

2. 统计邦健身俱乐部想进行一项调查，看看客户对他们的设施有何想法。

动动笔解答

考虑下面这些情况，你会选择什么作为目标总体？抽样单位是什么？你会如何拟定抽样空间？进行抽样时还需要考虑哪些问题？

1. 巧口华公司生产巧克力，他们为节庆季度限量生产了一些巧克力，想要检验这些巧克力的品质。

目标总体是全部限量版巧克力。

抽样单位是一块巧克力。

抽样空间需要涵盖所有巧克力，由于是限量生产，因此公司有可能记录生产了多少巧克力，包括每一种巧克力的数量。

在形成样本时，需要确保样本能代表总体，且不存在偏倚。如果这一批限量产品包含多种类型的巧克力，则要确保样本中包含每一类巧克力。

2. 统计邦健身俱乐部想进行一项调查，看看客户对他们的设施有何想法。

目标总体是统计邦健身俱乐部的所有客户。

抽样单位是一位客户。

抽样空间需要涵盖所有客户，有可能俱乐部有客户花名册，可以将这份花名册作为抽样空间。

和前面一样，你需要确保样本能够代表总体且没有偏倚，即确保客户性别、年龄等等都能得到全面的体现。

破案：消失的咖啡销量

咖啡为什么销量不佳?



我们无法肯定，但很有可能是因为实习生所调查的样本人群并未代表目标总体。

首先，实习生希望向友好客户免费派发咖啡，而且希望听到正面回应。这是不是说他只与看上去对他友好的客户交谈？他是得到了客户关于咖啡的真实评价，还是仅仅曾经询问他们是否同意“咖啡味道不错”？

实习生还希望利用这个工作机会和他常驻的咖啡店的一位年轻女常客搭讪，他是不是把大部分时间都花在这家店里了？这位女孩是否影响了他的样本选择？

最后，首席执行官推出咖啡的季节与进行调查的季节不同，这也有可能影响销量。所有这些因素，或者其中的部分因素，都有可能导致样本有误，进而导致了错误决策。

如何选择样本

我们已经讲过如何设计样本，也已经讨论过需要避免哪几类偏倚，现在我们需要从样本空间中选取实际样本，该怎么选呢？

简单随机抽样

一种做法是随机选取样本。假设你有一个包含N个抽样单位的总体，需要选取包含n个抽样单位的样本。**简单随机抽样**就是通过随机过程选取一个大小为n的样　本，所有大小为n的可能样本被选中的可能性都相同。

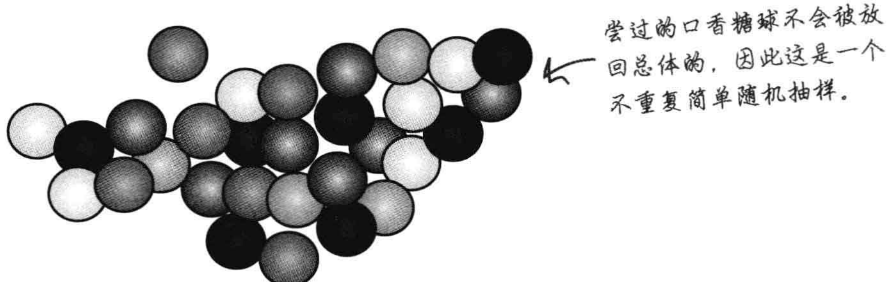
简单随机抽样有两种具体做法：**重复抽样**和**不重复抽样**。

重复抽样

重复抽样指的是：在选取一个抽样单位并记录下这个抽样单位的相关信息之后，再将这个单位放回总体中。这样做的结果是某个抽样单位有可能被选取不止一次。重复抽样的例子有：决定向大街上的行人提问，事前并不查看是否已经向该行人提问过。当你拦住行人请他们回答问题，然后在问完后让他们离开，实际上就是将行人放回了总体，这意味着你有可能不止一次向他们提问。

不重复抽样

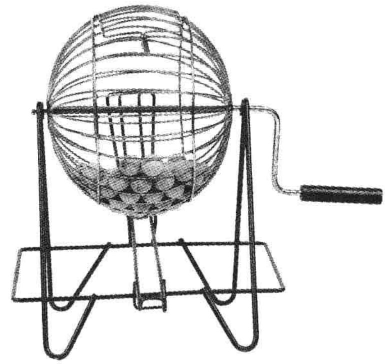
不重复抽样指的是：不再将抽样单位放回总体。不重复抽样的例子有：口香糖球检验—— 尝过的口香糖球是不会被放回总体的。



如何选取简单随机样本

使用简单随机抽样主要有两种方法：抽签，或使用随机编号。

抽签



抽签就是把抽样空间中的成员的名字或编号写在纸上或是球上，然后将其全部放入一个容器，再随机取出n个名字或编号，以便得到足够的样本单位。

随机编号生成器



如果你所处理的是一个大型抽样空间，抽签可能不太可行，于是可以采用另一种做法 —— 随机编号生成器或者随机编号表。这时，你为抽样空间的每个成员编一个编号，再生成一组共n个随机编号，然后从该空间中取出编号等于所生成的随机编号的成员。

重要提示：确保每个编号的生成机会相同，从而避免偏倚。

动动脑

简单随机抽样并不是不会发生问题，你觉得会在哪里出错？

其他类型的抽样

即便是简单随机抽样也免不了有问题。

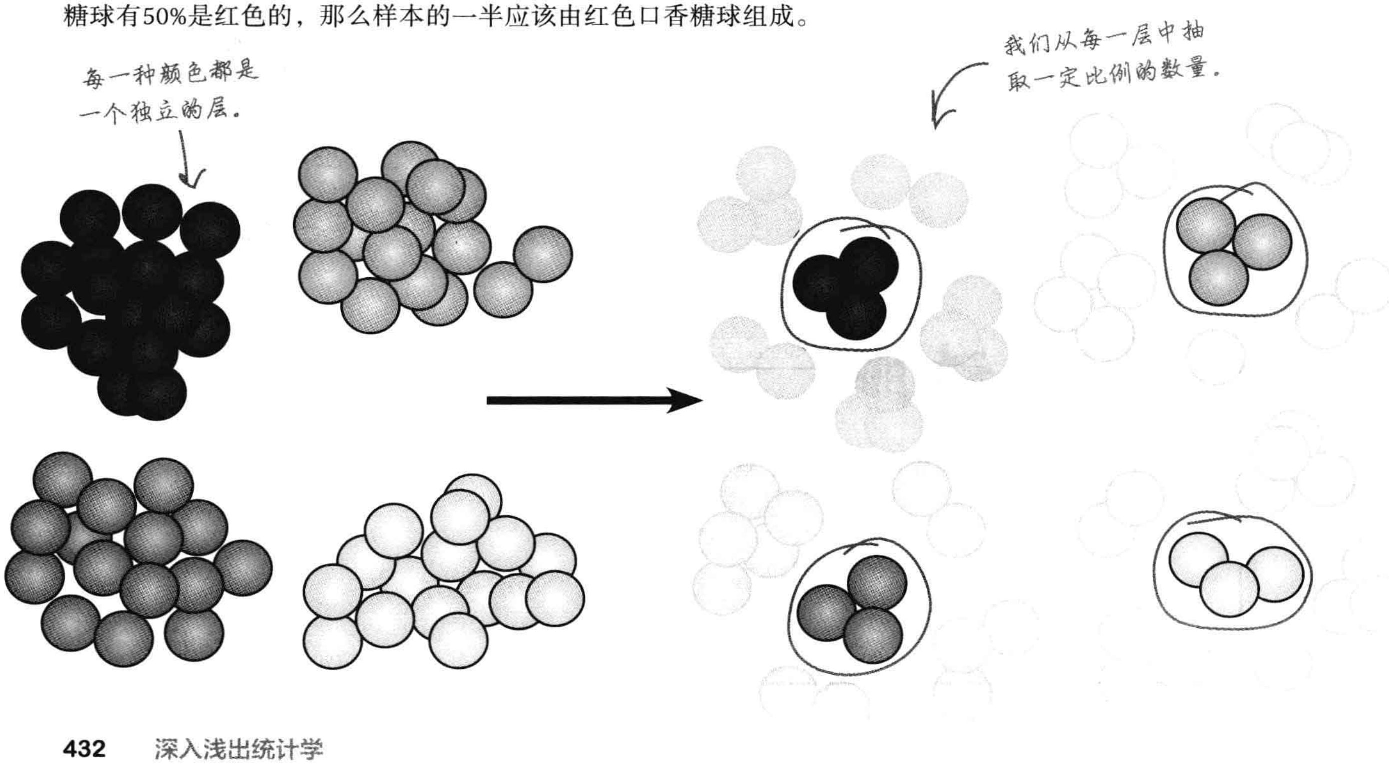
使用简单随机抽样时，仍然存在样本无法代表总体的可能性，例如，可能你最终随机抽到的全是黄色口香糖球，却错失其他颜色。

怎么避免这种情况呢？

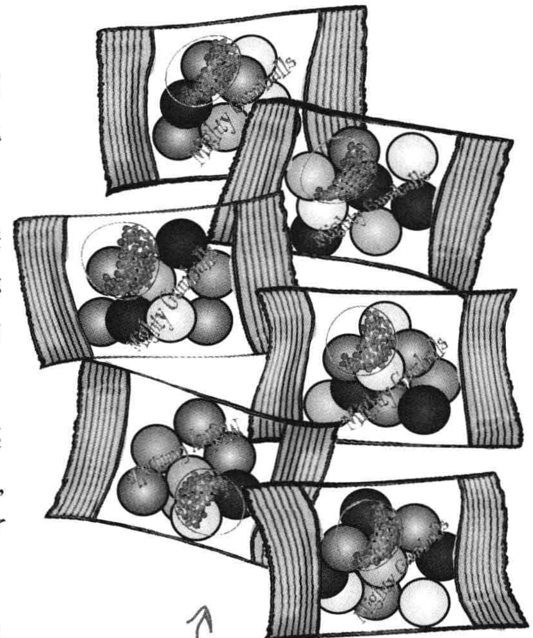
我们可以用分层抽样……

有一个方法可以取代简单随机抽样，即**分层抽样**。这种抽样类型将总体分割为几个相似的组，每个组具有类似的特性。这些特性或者组被称为**层**。例如，我们可以将口香糖球划分为不同的颜色——黄色、绿色、红色及粉色，这样每一种颜色就形成一个不同的层。

完成以上分层工作之后，就可以对每一个层进行简单随机抽样，确保最终样本中具有每一个组的代表。为此需查看每一个层在总体中所占的比例，然后按照相应比例从每一个层中抽取抽样单位。例如，如果曼帝糖果公司所生产的口香糖球有50%是红色的，那么样本的一半应该由红色口香糖球组成。



或可用整群抽样……



如果总体中包括大量相似的组或群，则**整群抽样**可以派上用场。例如，口香糖球可能会按盒出售，每一盒中的口香糖球的数量和颜色组成都相似，于是每一盒糖球形成一个**群**。

进行整群抽样时，不是对抽样单位进行简单随机抽样，而是**对群进行简单随机抽样**，然后对每一个群的各种特性进行调查。例如，你可以对一盒盒口香糖球进行简单随机抽样，然后品尝这些盒子里的糖球的味道。

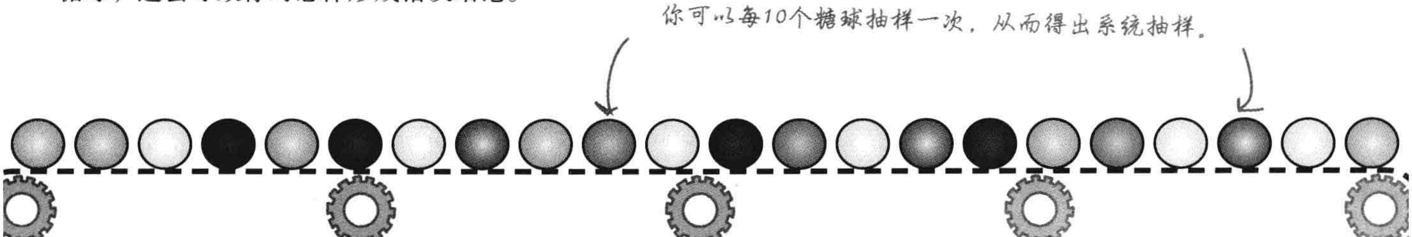
整群抽样之所以行得通，是因为群与群相互之间很相似，另外它还有一个优点，不需要使用总体抽样空间就可以进行整群抽样。例如，如果你正在调查树木情况，并把几片特定的森林作为群，则只需要了解你所选定的几片森林中的树木就行了。

整群取样的问题在于可能做不到完全随机。例如，很有可能一盒包装中的所有口香糖球都是同一个厂家生产的 —— 如果有不同生产厂家，你就不能选取这些糖球。

或甚至可用系统抽样

使用系统抽样时，按照某种顺序列出总体名单，然后每k个单位进行一次调查，其中k为一个特定数字。例如，可以选择每10个样本抽样一次。

相对而言，系统抽样既快捷又简单，但却有一个重大缺陷：如果总体中存在某种循环模式，则样本将会有偏倚。例如，如果糖球的生产工艺造成每到第10个糖球就是红色，那么你最终抽到的都是红色糖球，这会导致你对总体形成错误结论。



世上没有傻问题

问： 使用这些抽样方法能保证样本不存在偏倚吗？

答： 这些方法无法保证样本不存在偏倚，但能将发生偏倚的几率降至最低程度。通过认真思考目标总体，通过认真思考如何使样本代表总体，会更有机会得到无偏的代表性样本。

问： 我必须使用这些方法吗？不能随机选择对象吗？

答： 随机选择对象就是简单随机抽样。没错，你的确可以采用这种方法，但有一点要小心：你的样本有可能无法代表总体。

问： 可这是为什么？要是我随机选择对象，它们肯定会代表目标总体的。

答： 不一定。你看，如果你随机选择抽样单位，可能会选出一个无法有效代表目标总体的样本——这纯属随机现象。例如，如果完全随机地选择统计邦健身俱乐部的客户，有可能选出的都是同一个班的学员，或者选出的都是同一种性别的学员。还会出现这样的情况——你觉得自己是在随机选择，但事实并非如此。例如，如果你在进行顾客满意度调查，但却任凭顾客自己决定是否回应调查，那么，鉴于顾客只有得到充分鼓励才会给出回应，你最终得到的可能是有偏样本。最积极参与调查的顾客会是那些最满意或最不满意的顾客，那些感受不强烈的顾客发表意见的可能性则较小，然而，可能正是这部分人构成了总体的主要部分。

问： 要是我增大样本呢？能避免偏倚吗？

答： 样本越大，样本发生偏倚的几率越小，使用简单随机抽样时，这的确是一种使样本偏倚几率最小化的方法，问题在于样本越大，采集数据所需要的时间越多，工作越繁杂。

问： 分层抽样和整群抽样有何区别？

答： 分层抽样将总体划分为不同的组，或者叫做层，每一个层中的所有抽样单位相互之间尽量相似，也就是说，你认定一些特征或属性，例如性别，将其作为分层的依据。一旦将总体划分为层，就能对每个层进行简单随机抽样。整群抽样的目的是将总体划分为多个群，同时尽量保证群与群相似，随后通过简单随机抽样选取群，再接着就是对这些群中的对象进行抽样。

问： 明白了。这么说，在进行分层抽样时，要尽可能让每一个层不一样；而在进行整群抽样时，要尽可能让每一个群相似。

答： 完全正确。

问： 系统抽样怎么讲？

答： 进行系统抽样时，先选取一个数字k，然后，每到第k个对象就进行抽取，组成样本。这种抽样方法相当便捷，但这并不是说你的抽样一定可以代表总体。实际上，只有在抽样空间中不存在重复模式或组织时，这种抽样方式才能有效使用。

问： 抽签听起来很老套，大家仍在这么做吗？

答： 不如过去那样常用了，不过仍然是一种抽样方法。



有人给了你10盒巧克力，要求你对盒子里的巧克力进行抽样，盒子里有白巧克力、牛奶巧克力和黑巧克力。你的目标总体是所有巧克力，抽样单位是一块巧克力。

1. 如何用简单随机抽样解决这个问题?

2. 如何用分层抽样解决这个问题?

3. 如何用整群抽样解决这个问题?



练习解答

有人给了你10盒巧克力，要求你对盒子里的巧克力进行抽样，盒子里有白巧克力、牛奶巧克力和黑巧克力。你的目标总体是所有巧克力，抽样单位是一块巧克力。

1. 如何用简单随机抽样解决这个问题?

简单随机抽样：随机选取巧克力，可以用抽签方式，也可以用随机编号方式，如此一来，每一块巧克力都有同等的抽中机会。

2. 如何用分层抽样解决这个问题?

分层抽样：将巧克力分为不同的层，然后对每一层进行简单随机抽样。每一层都由特性相同的巧克力组成，因此可以按照巧克力的不同类型进行分层，可以将白巧克力作为一层，牛奶巧克力作为一层，黑巧克力作为最后一层。

3. 如何用整群抽样解决这个问题?

整群抽样：将巧克力分为几组，每一组都必须相似。假定每一盒巧克力都相似，则可以取其中一盒，然后对这一盒中的所有巧克力进行抽样。

练习

你会如何对曼帝糖果公司的超长效口香糖球进行抽样调查？糖球有4种颜色，都由同一家工厂生产。

假定你必须从零开始进行抽样。

练习解答

你会如何对曼帝糖果公司的超长效口香糖球进行抽样调查？糖球有4种颜色，都由同一家工厂生产。

假定你必须从零开始进行抽样。

目标总体是曼帝糖果公司的全部超长效口香糖球，抽样单位是单粒糖球，至于抽样空间，理想的情况是编制一份按编号排列的糖球表，但这可能无法付诸实现。因此我们用另一个方法来代替，即列出一个表，表中说明总体中的每种颜色的糖球各有多少粒。

使用何种抽样类型取决于你的主观意愿，但我们愿意选择分层抽样，因为这可能是得出无偏样本的最好方法。我们会将糖球按颜色进行划分，然后进行简单随机抽样，从四种颜色中选出一定比例的糖球，然后用这些糖球形成样本。

若你用了其他解决方法也无需担心，关键是想明白如何让你的调查最好地代表总体，具体办法可以不同。

要点

- **总体**是你所研究的所有事件的集合。
- **样本**是从总体中选取的相对较小的集合，可用于做出关于总体本身的结论。
- 进行抽样时，首先定义目标总体，即要研究的总体。然后确定抽样单位，即要抽样的对象类型。最后，拟定一个抽样空间，即目标总体中的所有抽样单位的列表。
- 如果样本不能代表目标总体，则这个样本存在**偏倚**。
- **简单随机抽样**即随机选择抽样单位并形成样本，包括重复抽样和不重复抽样。简单随机抽样的具体方式包括抽签或使用随机编号生成器。
- **分层抽样**即将总体划分为几个组，或者叫做几个层，组或层中的单位都很相似，每一层都尽可能与其他层不一样。分好层以后，就对每一层执行简单随机抽样。
- **整群抽样**即将总体划分为几个群，其中每个群都尽量与其他群相似，可通过简单随机抽样抽取几个群，然后用这些群中的每一个抽样单位形成样本。
- **系统抽样**即选取一个数字k，然后每到第k个抽样单位就抽样一次。

曼帝糖果公司有了样本

在你的帮助下，曼帝糖果公司采集到了超长效口香糖球的样本，这意味着不用尝遍整个糖球总体，而是用样本就可以进行检验了。



那么接下来做什么？

我们已经讲过如何采集具有代表性的样本，可还没有谈到如何利用这个样本。我们知道，一个无偏样本与总体具有相同的特征，但用哪种方法分析样本最好呢？

请接着往下读，下一章将讲解具体做法。

11 总体和样本的估计

进行预测



得样本而知总体，不亦乐乎？

若想成为**样本专家**，首先要懂得如何最有效地利用到手的样本——利用样本**准确地预测**总体，并以一定方式说明预测**结果的可靠程度**。在本章中，我们将讲解如何**通过样本了解总体**，以及如何通过总体了解样本。

糖球口味到底能持续多久？

在你的帮助下，曼帝糖果公司得到了超长效口香糖球的无偏样本，他们对样本中的每一粒糖球进行测试，得到了关于样本糖球口味持续时间的大量数据。

只有一个问题……

我不管样本的口味持续时间有多久，我只管总体的口味持续时间有多长，那样我才能宣布我们的糖球比别家的糖球嚼得久。

曼帝糖球公司首席执行官摩拳擦掌。



为了让首席执行官满意，我们需要求出曼帝糖果公司糖球总体的口味持续时间的均值和方差。

下面是我们从样本采集到的数据，你觉得我们该如何通过这些数据得出总体均值？

这是糖球口味持续时间，单位：分钟。

61.9	**62.6**	**63.3**	**64.8**	**65.1**
66.4	**67.1**	**67.2**	**68.7**	**69.9**

动动脑

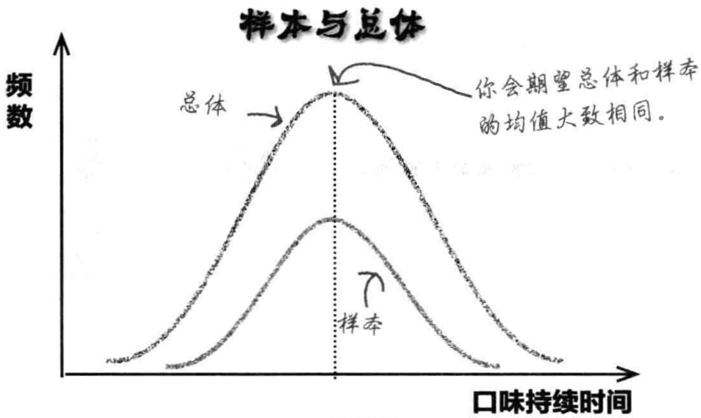
查看数据，你会如何使用这些数据估计总体的均值和方差？你觉得估计结果的可靠程度如何？为什么？

让我们首先估计总体均值

我们如何用糖球样本的口味测试结果得出糖球总体的口味持续时间均值？

答案其实十分直观。我们假设样本糖球的口味持续时间与总体糖球的口味持续时间相符，也就是说，我们求出样本的均值，然后将样本均值作为总体均值。

下面这张图显示了样本的分布情况以及可以基于样本而期望的总体分布情况。你会期望总体的分布与样本的分布相似，那样就能假设样本均值数值和总体均值数值大致相同。







不能说这二者完全吻合，但这是我们能做出的最好估计。

根据已知的情况，样本均值是我们能为总体均值做出的最好估计——在我们根据手头信息得到的数值中，样本均值是最有可能被作为总体均值的数值。

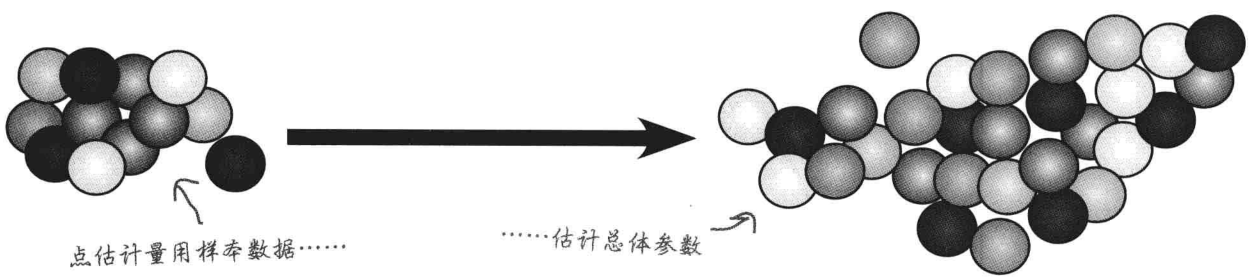
样本均值被称为总体均值的**点估计量**，也就是说，作为一个基于样本数据的计算结果，它给出了总体均值的良好估计。

点估计量可以近似总体参数

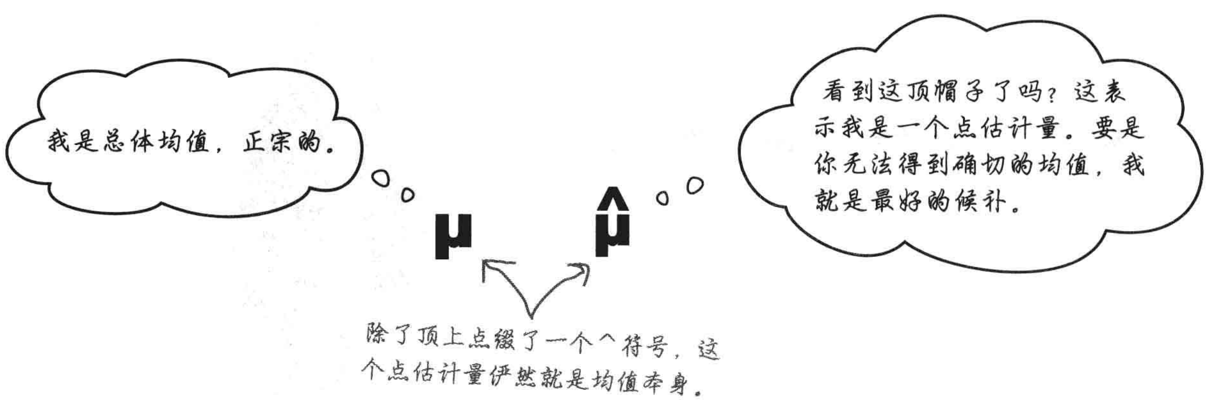
在此之前，我们用到过一些总体参数的实际值，如均值μ，或方差σ^2。我们要么能够自己动手算出这些数值，要么已经知道这些数值是多少。

而这一次，我们不知道总体参数的确切数值。我们无法通过总体**计算**这些参数，而只能通过样本数据**估计**这些参数。于是，我们用“点估计量”对总体参数进行最接近的猜测。

一个总体参数的**点估计量**就是可用于估计总体参数数值的某个函数或算式，例如，由于我们能用样本均值估计总体均值，因此样本均值就是总体均值的点估计量。



我们用符号^区别实际总体参数和它的点估计量，例如：用符号μ表示总体均值，而用$\hat{\mu}$表示样本均值，即，为了指出你正在使用的是某一个总体参数的点估计量，则在该总体参数的符号上方标上^。





样本均值有一个简记符。

符号μ具有十分精确的含义——总体的均值。为了不至于混淆，样本均值另有一种表示方法：$\bar{x}$（读作“x拔”）。这样一来，当别人说到μ时，我们就知道是指总体均值；说到$\bar{x}$时，就知道是指样本均值。

$\bar{x}$是μ的样本对等量，它的计算方法和总体均值的计算方法一样——将样本中的所有数据加起来，除以总数。即，如果样本大小为n，则：

$$\bar{x} = \frac{\Sigma x}{n}$$

$\bar{x}$ 是样本的均值。

将样本中的数字相加，然后除以这些数字的总数。

我们可以根据上式写出总体的点估计量的简明表达式，由于可以用样本均值估计总体均值，因此：

$$\hat{\mu} = \bar{x}$$

我们估计总体均值…… ……用的是样本均值

动动笔

使用样本数据估计总体均值的数值。提示数据如下：

61.9　62.6　63.3　64.8　65.1　66.4　67.1　67.2　68.7　69.9

动动笔解答

使用样本数据估计总体均值的数值。提示数据如下：

61.9　62.6　63.3　64.8　65.1　66.4　67.1　67.2　68.7　69.9

我们可以通过计算样本均值估计总体均值：

$$\hat{\mu} = \bar{x} = \frac{61.9 + 62.6 + 63.3 + 64.8 + 65.1 + 66.4 + 67.1 + 67.2 + 68.7 + 69.9}{10}$$

$$= 657/10$$

$$= 65.7$$

世上没有傻问题

问： 均值就是均值，怎么用这么多符号来表示？

答： 用到的概念有三个：总体均值、样本均值以及总体均值的点估计量。

总体均值用 μ 表示，本书前面一直在讲的就是这种均值，其计算方法是：将总体中的所有数据相加，然后除以数据个数之和。

样本均值用 $\bar{x}$ 表示，计算方法同 μ，不过这时用的是样本中的数据。$\bar{x}$ 的算法是：将样本中的所有数据相加，然后除以样本个数之和。

点估计量用 $\hat{\mu}$ 表示，它其实是根据样本数据得出的对你所认为的总体均值的最佳猜测值。

问： 这是不是意味着我们只要算出样本均值就能求出 μ？

答： 我们无法通过样本求出 μ 的确切数值，不过，只要样本是无偏的，就能得出十分接近的估计值。即，我们可以利用样本数据求出 $\hat{\mu}$，但无法求出 μ 本身的真值。

问： 如果样本是有偏的会怎么样？如何计算 μ 的估计值？

答： 尽量让样本无偏的重要性就体现在这里。如果你手头的数据都来自样本，那么就要将样本作为估计基础。如果样本有偏，就意味着 μ 的估计值有可能不准确，有可能因此做出错误的估计。

问： 样本的大小有影响吗？

答： 一般说来，样本越大，点估计量越准确。

μ 是总体均值，$\bar{x}$ 是样本均值，$\hat{\mu}$ 是 μ 的点估计量。

要点

- **点估计量**由样本数据得出，是对总体参数的估计。
- 在讨论总体参数的点估计量时，会为总体参数添上一个^符号。例如 μ 的点估计量写作 $\hat{\mu}$。
- 计量样本的均值用 $\bar{x}$ 表示，样本的均值可用下列公式进行计算：

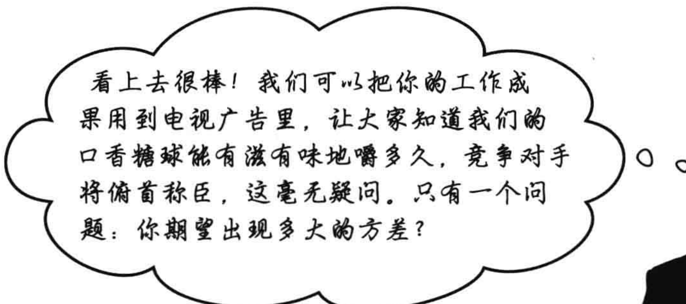
$$\bar{x} = \frac{\Sigma x}{n}$$

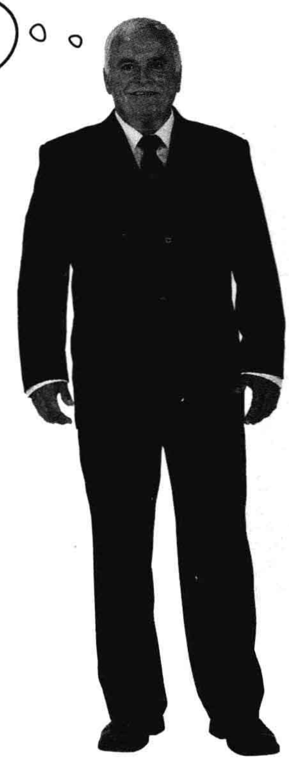
其中x代表各个样本的数值，n为样本的个数。

- 通过计算 $\bar{x}$ 可得到总体均值的点估计量，即：

$$\hat{\mu} = \bar{x}$$

这说明，如果想十分近似地估计总体均值的真值，可以使用样本均值。





你已经得到了总体均值的良好估计，那么方差呢？

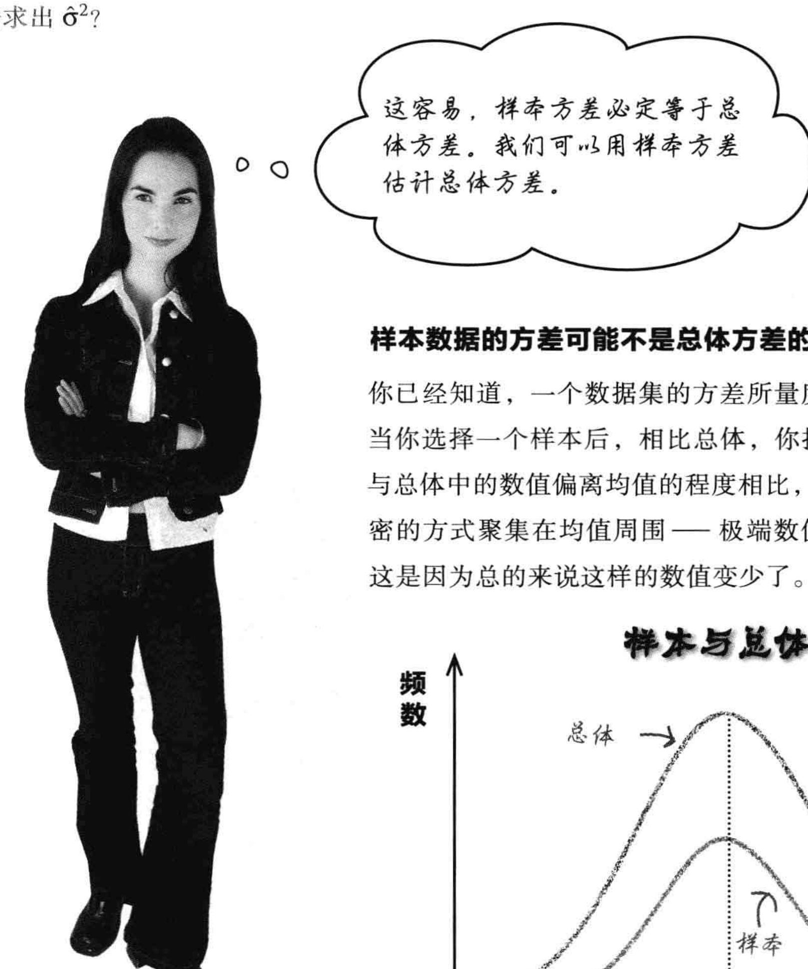
只要我们得出总体方差的良好估计，首席执行官就能根据样本数据的结果，判断糖球总体的口味持续时间有可能出现多大变异。

让我们估计总体方差

前面讲到如何利用样本均值估计总体均值，也就是说，我们为超长效口香糖球总体找到了一个估计口味持续时间均值的办法。

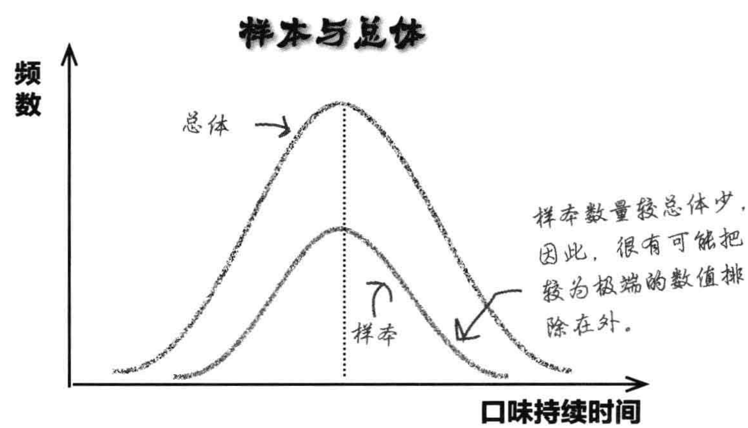
为了让曼帝糖果公司首席执行官感到满意，我们还需要得出总体方差的良好估计。

我们可以把哪个算式当作总体方差的点估计量呢？即，我们该如何利用样本数据求出 $\hat{\sigma}^2$？



样本数据的方差可能不是总体方差的最好估计办法

你已经知道，一个数据集的方差所量度的是数值与均值的偏离程度。当你选择一个样本后，相比总体，你拥有的数值数量变少了，因此，与总体中的数值偏离均值的程度相比，样本中的数值更有可能以更紧密的方式聚集在均值周围 —— 极端数值出现在样本中的可能性下降，这是因为总的来说这样的数值变少了。



那么哪个算式能更好算出总体方差的估计值呢？

我们需要一个有别于样本方差的点估计量

用样本方差估计总体方差会出现这样的问题：估计结果会稍微偏低 —— 样本方差可能会略小于总体方差，差别程度则取决于样本数值的大小。样本较小时，样本方差与总体方差的差别有可能更大。

我们需要找到一个更好的办法来估计总体方差 —— 找到样本数据的某个函数，而这个函数所得出的结果要稍微大于所有样本数值的方差。

那么用哪个算式作为估计量？

我们不使用样本数据的方差，而用其他方式估计总体方差。如果样本大小为n，可以用下列算式估计总体方差：

$$\hat{\sigma}^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

总体方差估计量

用样本中的每一个数值减去样本均值，所得之差取平方数，然后将所有平方值相加。

除以样本大小减1。

即，取样本中的每一个数值，减去样本均值，所得之差取平方数；然后将所有平方值加起来，除以样本数减1。这个算法与样本方差的算法相似，不过除数是n－1，而不是n。



这个公式与总体方差的数值更接近。

一组数字除以n－1的结果大于这一组数字除以n的结果，当n相当小时，这种差别最为显著。也就是说，通过公式算得的结果与样本数据的方差近似，但会略微偏大。

总体方差往往大于样本数据的方差，因此，用这个公式作为总体方差的点估计量，效果更好一点儿。

方差细细看

要想知道用哪个公式求方差，很需要费点思量。一个是求总体方差σ^2的公式，一个是略有变化的求总体方差点估计量$\hat{\sigma}^2$的公式，什么时候用这个？什么时候用那个？

求总体方差

如果想求确切的总体方差，且拥有全部总体数据，则可用下式进行计算：

$$\sigma^2 = \frac{\Sigma(x - \mu)^2}{n}$$

总体方差 → σ^2；总体均值 → μ；总体大小 → n

在这种情况下：你拥有所有总体数据；你知道总体均值；你想求出所有这些数值的方差——这正是前面一直在用的计算方法。

估计总体方差

如果需要用样本数据估计总体方差，则可用下式进行计算：

$$\hat{\sigma}^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

基于样本的总体方差点估计量 → $\hat{\sigma}^2$；样本均值 → $\bar{x}$

是n－1，而不是n，n是样本的大小。这里算的是估计值。

上式不是在“计算”有n个数值的实际总体的方差，而是根据所拥有的样本数据来“估计”总体方差。为了估计得更准确一些，除数用了n-1，而不是n，这样就能得出略大一点儿的结果。

总体方差点估计量的式子通常写作s^2，由此得到：

$$\hat{\sigma}^2 = s^2 \quad \text{其中} \quad s^2 = \frac{\Sigma(x - \bar{x})^2}{n - 1}$$

总体方差的点估计量 → $\hat{\sigma}^2$

s^2给出了基于样本数据的公式。

这种表示方法类似于用$\bar{x}$表示样本均值。

哪个公式用在哪里？

是用n做除数求方差，还是用n-1做除数求方差？这个问题有时候真是让人愁肠百结。做出判断的黄金准则是：用n做除数会得出“**手头拥有的一组数据的实际方差**”。

如果手头拥有整个总体的数据，则以n为除数会得出总体的实际方差——需要所用σ^2的公式，除数为n。

如果手头拥有总体的一个样本的数据，则你可能会希望用这个样本估计总体方差——需要使用s^2公式，除数为n-1。



有些书上说计算样本时用n−1，有些书则说用n。

这是因为每一本书对样本的用途作了不同的假设，如果要用样本估计总体方差，则要除以n−1。只有在需要计算一组确切数值的方差时，才除以n。

如果你正在参加统计学考试，请问清考试委员会指定的方法。

动动笔

下面是曼帝糖果的样本数据。
请你估计，总体方差是多少？

61.9　62.6　63.3　64.8　65.1　66.4　67.1　67.2　68.7　69.9

动动笔解答

下面是曼帝糖果样本的数据。
请你估计，总体方差是多少？

61.9 62.6 63.3 64.8 65.1 66.4 67.1 67.2 68.7 69.9

我们可以通过计算s^2估计总体方差。

$$\hat{\sigma}^2 = s^2$$

$$= \frac{\sum (x - \bar{x})^2}{n - 1}$$

$$= \frac{(-3.8)^2 + (-3.1)^2 + (-2.4)^2 + (-0.9)^2 + (-0.6)^2 + (0.7)^2 + (1.4)^2 + (1.5)^2 + (3)^2 + (4.2)^2}{9}$$

$$= \frac{14.44 + 9.61 + 5.76 + 0.81 + 0.36 + 0.49 + 1.96 + 2.25 + 9 + 17.64}{9}$$

$$= 62.32/9$$

$= 6.92$（保留两位小数）

世上没有傻问题

问： 为什么计算样本方差要除以n-1？为什么不能除以n？

答： 这是因为，在大部分情况下都是用样本数据估计总体方差。除以n-1比除以n能得出精确性稍微高一点儿的结果，因为样本数值的方差很可能略小于总体方差。

问： 这有数学依据吗？

答： 有啊，我们会在本章末尾讲到这一点。能想到这一点很不错，请继续保持。

问： 我该如何记住哪个符号用于总体，哪个符号用于样本？

答： 一般说来，希腊字母用于表示总体参数，而普通罗马字母用于表示样本的均值和方差。

问： 能像求方差的点估计量一样求出标准差的点估计量吗？怎么做？

答： 为了估计标准差，首先要计算方差的估计量，标准差的估计量等于方差估计量的平方根。

曼帝糖果公司抽取了更多样本

口味测试结果让曼帝糖果公司首席执行官大受鼓舞，他要求再进行一次抽样，以便发布电视广告。这一次，首席执行官希望能够宣传曼帝糖果公司的产品相比竞争对手的产品有多么受欢迎。

曼帝糖果公司的职员随机抽取了一些人，问他们是喜欢曼帝公司生产的口香糖球还是喜欢曼帝公司竞争对手生产的口香糖球。职员们希望能够利用调查结果预测：总体中有多大比例的人“可能偏爱曼帝公司的糖球”。





曼帝糖果公司发现，在40个人中有32个人偏爱他们的口香糖球，其余8个人则偏爱竞争对手的口香糖球。

动动脑

你会如何求出样本中偏爱曼帝糖果口香糖球的人所占的比例？你认为这符合哪种分布？如何将求得的结果用于总体？

这是一个比例问题

对于曼帝糖果的最新抽样，首席执行官感兴趣的是，是否人人都偏爱曼帝糖果的产品，而不是偏爱竞争对手的产品。也就是说，可以将偏爱曼帝糖果的每一个人作为一个“成功”事件。

那么我们如何利用样本数据预测总体的“成功”比例？

预测总体比例

如果我们用X表示总体的成功事件数量，则X符合二项分布，参数为n和p。n为总体中的人数，p为成功事件的比例。

就像总体均值的最接近估计是样本均值一样，总体成功比例的最接近猜测肯定是样本成功比例。即，如果我们求出样本中偏爱曼帝糖果的人的比例，就能十分近似地估计出总体人群中偏爱曼帝糖果的人的比例。

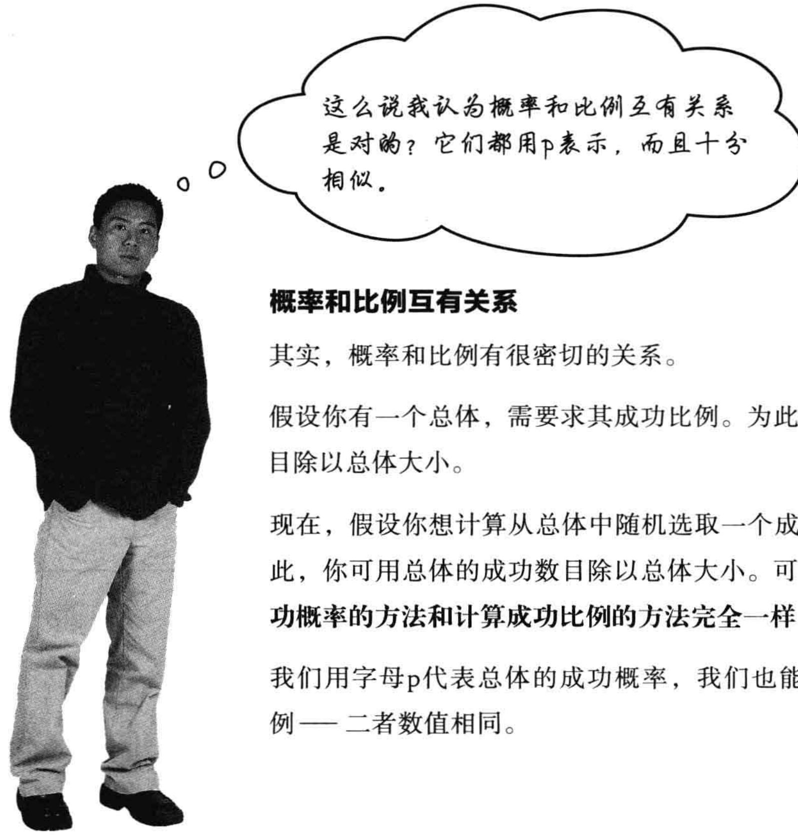
用偏爱曼帝糖果的总人数除以样本总人数，就能得出样本的成功比例；如果用p_s代表样本的成功比例，则可以下式估计总体的成功比例：

总体成功比例的点估计量 → $\hat{p} = p_s$ ← 样本成功比例

其中

$$p_s = \frac{\text{成功数目}}{\text{样本数目}}$$

也就是说，我们将样本成功比例作为总体成功比例的点估计量，在曼帝糖果的最新抽样中，40个人中有32个人偏爱曼帝糖果产品，因此$p_s = 0.8$。于是，总体成功比例的点估计量也是0.8。



概率和比例互有关系

其实，概率和比例有很密切的关系。

假设你有一个总体，需要求其成功比例。为此，你可用成功的数目除以总体大小。

现在，假设你想计算从总体中随机选取一个成功事件的概率。为此，你可用总体的成功数目除以总体大小。可以看出，你**计算成功概率的方法和计算成功比例的方法完全一样**。

我们用字母p代表总体的成功概率，我们也能方便地用p代表比例—— 二者数值相同。

p = probability（概率）= proportion（比例）

动动笔

曼帝糖果公司为超长效口香糖球取得了另一个样本，并发现，在样本中，40个人中有10个人偏爱粉色口香糖球，这些人对其他颜色不那么喜欢。总体中偏爱粉色糖球的人的比例是多少？从总体中选中一个不偏爱粉色糖球的人的概率是多少？

动动笔解答

曼帝糖果公司为超长效口香糖球取得了另一个样本，并发现，在样本中，40个人中有10个人偏爱粉色口香糖球，这些人对其他颜色不那么喜欢。总体中偏爱粉色糖球的人的比例是多少？从总体中选中一个不偏爱粉色糖球的人的概率是多少？

我们可以利用样本比例估计总体比例，即：

$$\hat{p} = p_s = 10/40$$
$$= 0.25$$

从总体中选中一个不喜欢粉色糖球的人的概率：

$$P(\text{不偏爱粉色}) = 1 - \hat{p}$$
$$= 1 - 0.25$$
$$= 0.75$$

世上没有傻问题

问： 这么说比例和概率是一回事？

答： 总体的成功数目除以总体大小即等于比例，这个算法和用于计算二项分布的概率的算法是一样的。

问： 比例算法仅适用于二项分布吗？是否适用于其他概率分布？

答： 在我们讲过的所有概率分布中，二项分布是唯一与比例有关的分布。比例算法专门用于解决这种分布的问题。

问： 样本比例与总体比例一样吗？

答： 样本比例可以作为总体比例的点估计量，其实，样本比例是对总体比例的具体数值的最好猜测。

问： 如果样本存在偏倚，还会这样吗？如何通过一个有偏样本估计比例？

答： 关键在于确保样本无偏，因为样本是估计的依据。如果样本有偏，那么就会得出对总体比例的不准确估计。其他点估计量也会有这种情况。

问： 那么如何确保样本无偏呢？

答： 请复习前面章节讲过的要点，遵守这些要点是确保样本尽量具有代表性的好办法。多花点力气准备样本是值得的，这意味着你的点估计量能够更精确地反映总体本身。



要 点

- 总体方差的点估计量如下：

$$\hat{\sigma}^2 = s^2$$

其中s^2的算法为：

$$\frac{\Sigma(x - \bar{x})^2}{n - 1}$$

- 总体比例用p表示，即总体的成功比例。

- P的点估计量为p_s，其中p_s为样本的成功比例。

$$\hat{p} = p_s$$

- p_s的计算方法是：用样本中的成功数目除以样本数目。

$$p_s = \frac{\text{成功数目}}{\text{样本数目}}$$

快来这儿买糖球！

还记得统计邦电影院吗？他们最近获得特许，可以销售曼帝糖果，这个动向证明很多顾客都喜欢曼帝糖果公司的糖球。

问题是，并非人人都开心。



引进大盒装糖球

电影院出售混合型盒装糖球；还有，这个周末他们将播出一系列经典老片。

这次活动看来很受欢迎，出票情况很好。问题是，有的人要是吃不到自己喜欢的红色糖球就会大失所望。

一盒大包装糖球可供数人分享，每一盒装有100粒糖球；糖球总体中有25%是红色的。



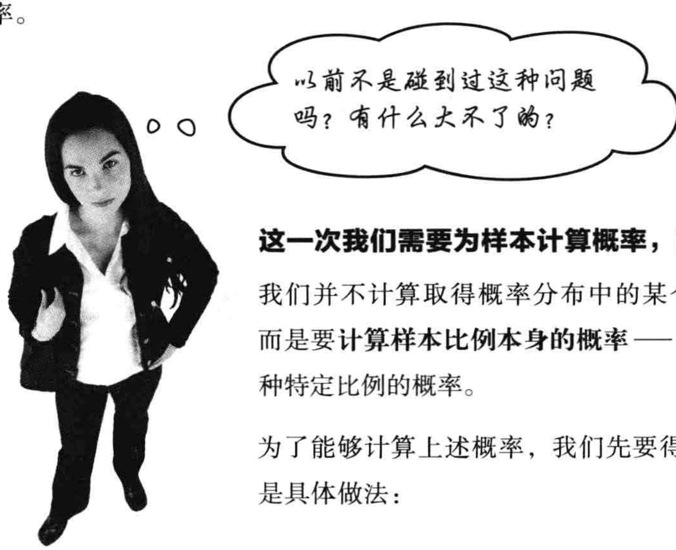
我们需要求一大盒特定糖球中有40颗或40颗以上红色糖球的概率。

由于每一大盒糖球的容量为100颗，也就是说我们要求出在一大盒特定糖球中红色糖球占40%的概率，且已知糖球总体的25%是红色的。

这和抽样有什么关系？

前面已经讲过如何得到无偏样本，以及如何利用样本求出总体参数的点估计量。

这一次，情况有所不同——总体参数已知，需要为某一盒特定糖球计算概率。也就是说，在这里要算的不是总体的概率，而是样本比例的概率。



这一次我们需要为样本计算概率，而不是为总体计算概率。

我们并不计算取得概率分布中的某个特定频数或特定数值的概率，而是要**计算样本比例本身的概率**——我们要算出在一个整体中出现一种特定比例的概率。

为了能够计算上述概率，我们先要得出样本比例的概率分布，下面是具体做法：

1. **查看与我们的特定样本大小相同的所有样本。**
   如果我们有一个大小为n的样本，就需要考虑所有大小为n的可能样本。在本例中，盒子里的糖球数量为100，因此n为100。

2. **观察所有样本比例形成的分布，然后求出比例的期望和方差。**
   每一个样本都有自己的情况，因此每个包装盒里的红色糖球的比例都有可能发生变化。

3. **得出上述比例的分布后，利用该分布求出概率。**
   得知一个样本中的“成功比例”的分布后，就能够利用这个分布求出一个随机样本的比例概率——这里的随机样本是一大盒糖球。

让我们看看具体做法。

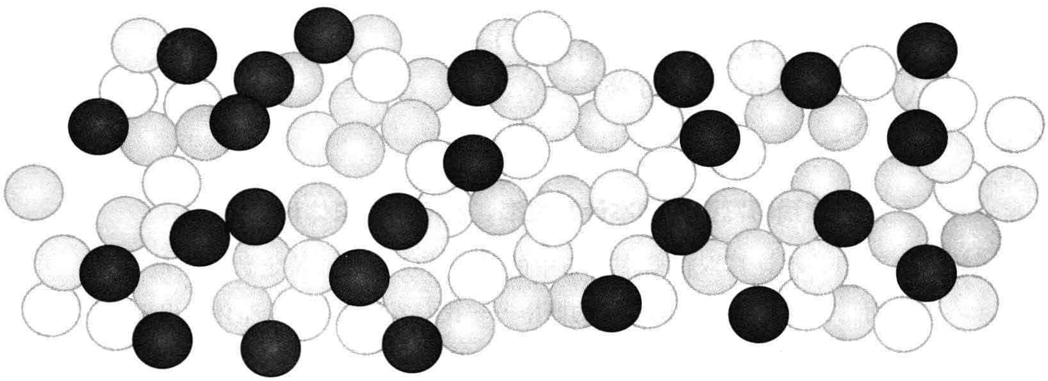
比例的抽样分布

如何求样本比例的分布？

让我们先看糖球总体。已知总体中的红色糖球的比例，用p表示，即p=0.25。

糖球总体

糖球总体的25%为红色，因此p = 0.25。



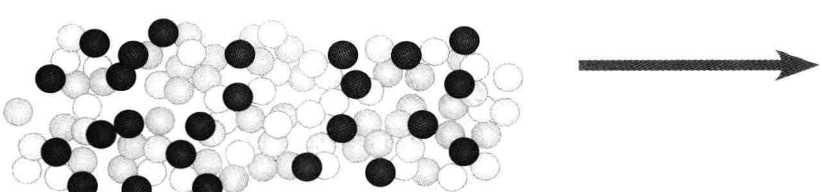
每一大盒糖球其实就是从糖球总体中取出的一个样本。每一大盒装有100颗糖球，因此样本大小为100，让我们用n表示这个大小。

如果用随机变量X代表样本中的红色糖球的数目，则X~B(n,p)，其中n=100，p = 0.25。

样本中的红色糖球的比例取决于X —— 样本中的红色糖球的数目，即比例本身是一个随机变量，可以将此记为P_s，且P_s = X/n。

样本

P_s表示样本的成功比例。



$$P_s = \frac{X}{n}$$

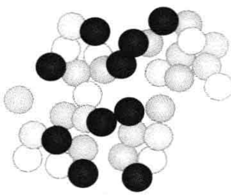
X ~ B(n, p)

我们不知道样本中的红色糖球的确切数目，但我们知道它的分布。

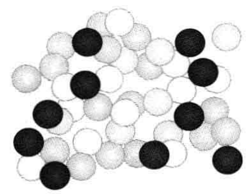
可以取出的大小为n的可能样本为数众多。每一个可能样本会包含n颗糖球，每一盒样本中的红色糖球的数量会符合相同的分布——对于每一个样本，红色糖球的数量符合B(n, p)，成功比例则为X/n。

和前面讲到的样本一样，每一个样本都包含n个元素。

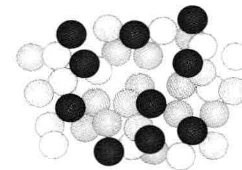
几个不同的样本



X ~ B(n, p)
P_s = X/n



X ~ B(n, p)
P_s = X/n



X ~ B(n, p)
P_s = X/n

利用所有可能的样本，我们能得出所有样本比例的分布，该分布称作“比例的抽样分布”，或者称作“P_s的分布”。

明白了。“比例的抽样分布”其实是一种概率分布，由所有大小为n的可能样本的各种比例构成。如果我们知道这些比例的分布，就能用这个分布求出某一个特定样本的比例的发生概率。



利用比例的抽样分布，能够求出某一个随机选择的、大小为n的样本的“成功比例”的概率。

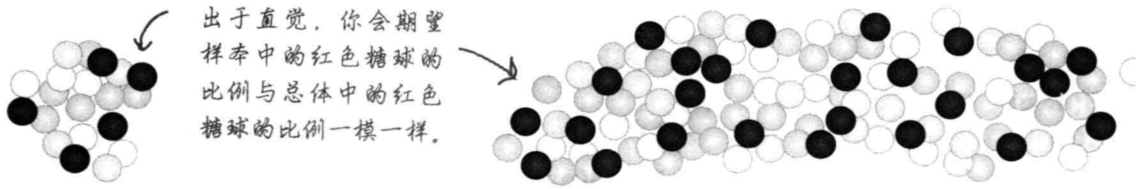
也就是说，我们能够利用比例的抽样分布求出“某一大盒糖球中的红色糖球比例至少为40%”的概率。

不过，在此之前，我们需要知道上述分布的期望和方差。

P_s的期望是多少？

前面讲到，我们可以通过所有可能取用的、大小为n的样本的各个比例形成一个分布，为了能够用这个分布计算概率，我们还需要了解更多有关这个分布的数据——尤其需要知道方差和期望。

让我们先考虑期望。出于直觉，我们会期望样本中的红色糖球的比例与总体中的红色糖球的比例保持一致。如果糖球总体中包含25%的红色糖球，那么，可以期望样本中也包含25%的红色糖球。



那么P_s的期望是多少？

我们想求E(Ps)，其中 $P_s = X/n$。也就是说，我们想求出所期望的样本比例数值，这里的样本比例等于红色糖球的数量除以样本糖球的总数量，即：

$$E(P_s) = E\left(\frac{X}{n}\right)$$

$$= \frac{E(X)}{n}$$

上式中的X为样本中的红色糖球的数目，如果我们把红色糖球数目视为“成功数目”，则$X \sim B(n, p)$。

在二项分布一章已经讲过：E(X) = np，于是：

$$E(P_s) = \frac{E(X)}{n}$$

$$= \frac{\cancel{n}p}{\cancel{n}} \quad \leftarrow E(X) = np$$

$$= p$$

这个结果正好符合我们直觉中的期望。我们可以期望样本的成功比例与总体的成功比例相一致。

P_s的方差是多少？

为了能够求出任何样本比例的概率，我们还需要先知道P_s的方差——可以用求期望的相似方法求方差。

那么$Var(P_s)$是多少？让我们像以前一样，从$P_s = X/n$开始：

$$Var(P_s) = Var\left(\frac{X}{n}\right)$$

$$= \frac{Var(X)}{n^2}$$

来自于$Var(aX) = a^2Var(X)$，在本例中，$a = 1/n$。

如上所述，X为样本中的红色糖球的数目。如果我们将红色糖球的数目视为“成功数目”，则$X \sim B(n, p)$，于是$Var(X) = npq$，即二项分布的方差。得到：

$$Var(P_s) = \frac{Var(X)}{n^2}$$

$$= \frac{npq}{n^2}$$

$Var(X) = npq$

$$= \frac{pq}{n}$$

取方差的平方根，可得P_s的标准差，它指出样本比例与p的可能差距，有时候我们将它称为“比例标准误差”，因为它能指出样本比例的可能误差。

$$\textbf{比例标准误差} = \sqrt{\frac{pq}{n}}$$

n越大，比例标准误差越小。也就是说，样本中包含的对象越多，用样本比例作为p的估计量就越可靠。

现在，如何用所求得的期望和方差数值计算比例的概率呢？让我们接着进行下去。

求解P_s的分布

我们在前面求出了P_s—— 比例的抽样分布的期望和方差，我们发现，如果通过所有样本比例形成一个分布，则：

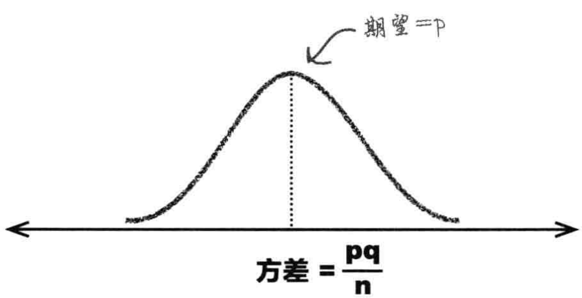
$$E(P_s) = p \qquad Var(P_s) = \frac{pq}{n}$$

我们可以借助以上结果求出“大小为100的样本中的红色糖球的比例至少为40%”的概率。



没错，P_s的分布实际上取决于样本的大小。

下面是一张P_s的分布图，其中n很大。



动动脑

观察P_s的分布图形，这里n很大。你觉得P_s符合什么分布？

P_s符合正态分布

当n很大时，P_s的分布接近正态分布。所谓“很大”是指大于30。n越大，P_s的分布越接近正态分布。

我们已经求得P_s的期望和方差，也就是说，当n很大的时候：

$$P_s \sim N\left(p, \frac{pq}{n}\right)$$

由于在n>30的时候P_s符合正态分布，所以可以用正态分布解答我们的糖球问题。我们可以用正态分布计算“某一大盒糖球中的红色糖球比例至少为40%”的概率。

只是有一件事别忘了：需要对抽样分布进行连续性修正。



有时候统计学家对n应该为多大无法达成共识。

如果你正准备参加统计学考试，一定要问清楚考试委员会的要求。

P_s—需要进行连续性修正

每个样本的“成功数目”都是离散的。由于使用“成功数目”计算比例，因此在用正态分布计算概率时，要进行连续性修正。

我们前面讲过，如果用X表示样本中的成功数目，则$P_s = X/n$；X的正态连续性修正为 ± (1/2)。

如果我们用以上数值替代公式$P_s = X/n$中的X，那么P_s的连续性修正为：

$$\text{连续性修正} = \frac{\pm(1/2)}{n} = \frac{\pm 1}{2n}$$

即，如果用正态分布近似计算P_s的概率，一定要用 ± 1/2n进行连续性修正；连续性修正的确切数值取决于数值n。



放轻松

如果n很大，则可以忽略连续性修正。

随着n增大，连续性修正变得很小，于是对整个概率带来的变化极小。有些课本会完全忽略连续性修正。

世上没有傻问题

问： 什么是抽样分布?

答： 如果从一个总体中用相同的方法抽取许多大小相同但存在差异的样本，然后用每个样本的某个属性形成一个分布，则所得结果称为抽样分布。由此得出，用每个样本的比例形成的抽样分布就是“比例的抽样分布”。

问： 我们的确需要采集所有可能采集的样本吗?

答： 不，其实我们不用实际动手采集所有样本，而是假设我们采集了所有样本，然后得出期望和方差的表达式。

问： 这么说抽样分布有期望和方差? 为什么?

答： 抽样分布是一个概率分布，因此，像其他概率分布一样，它有期望和方差。

比例的抽样分布的期望类似于样本比例的平均数，等于从一个特定总体中取出的样本的期望比例。

问： 为什么P_s的方差和总体方差σ^2不一样?

答： 比例的抽样分布的方差描述的是样本比例的变化情况，而不是描述数值本身的变化情况。由于描述的概念不一样，因此结果数值不一样。

问： 比例的抽样分布有什么用处?

答： 可以用它求出从一个已知总体中取出的某个样本的比例的概率，可以由此得知样本的期望形态。

问： 比例标准误差究竟有何含义?

答： 标准误差是抽样分布的方差的平方根，实际上，它指出你能够期望的样本比例与总体比例真值的差距，即指出你能期望出现哪种误差。

要 点

- 考虑从同一个总体中取得的所有大小为n的可能样本，由这些样本的比例形成一个分布，这就是“比例的抽样分布”。我们用P_s代表样本比例随机变量。

- Ps的期望和方差的定义式是:

$$E(P_s) = p$$

$$Var(P_s) = pq/n$$

其中p为总体比例。

该分布的标准差称为**比例标准误差**，其定义式为:

$$\sqrt{Var(P_s)}$$

- 如果n > 30，则Ps符合正态分布，于是:

$$P_s \sim N(p, pq/n)$$

使用这个公式时需要进行连续性修正:

$$\pm\frac{1}{2n}$$



练习

糖球总体的25%为红色。在一盒装有100粒糖球的包装盒中，至少有40%红色糖球的概率有多大？让我们逐步进行计算。

1. 如果P_s表示盒中的红色糖球的比例，那么P_s符合什么分布？

2. $P(P_s \geqslant 0.4)$的数值是多少？

提示：别忘了进行连续性修正。



糖球总体的25%为红色。在一盒装有100粒糖球的包装盒中，至少有40%红色糖球的概率有多大？让我们逐步进行计算。

1. 如果P_s表示盒中的红色糖球的比例，那么P_s符合什么分布？

让我们用p表示盒中红色糖球的概率。即$p = 0.25$。

让我们用P_s表示盒中红色糖球的比例。

$P_s \sim N(p, pq/n)$，其中$p = 0.25$，$q = 0.75$，且$n = 100$。由于pq/n等于$0.25 \times 0.75 / 100 = 0.001875$，得到：

$P_s \sim N(0.25, 0.001875)$

2. $P(P_s \geq 0.4)$的数值是多少？提示：别忘了进行连续性修正。

$$P(P_s \geq 0.4) = P(P_s > 0.4 - 1/(2 \times 100))$$
$$= P(P_s > 0.395)$$

由于$P_s \sim N(0.25, 0.001875)$，我们需要求出0.395的标准分，这样就能在概率表中查找结果。于是得出：

$$z = \frac{0.395 - 0.25}{\sqrt{0.001875}}$$
$$= 3.35$$

$$P(Z > z) = 1 - P(Z < 3.35)$$
$$= 1 - 0.9996$$
$$= 0.0004$$

即一盒100粒装的糖球中，红色糖球至少为40%的概率是0.0004。



比例的抽样分布细细看

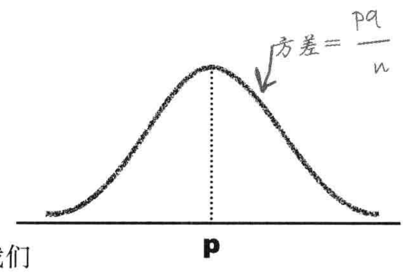
取所有大小为n的可能样本的比例，形成分布，这就是比例的抽样分布。一个样本的成功比例用P_s表示，且：

$$E(P_s) = p$$

$$Var(P_s) = \frac{pq}{n}$$

当n很大时，例如大于30，则P_s近似为正态分布，于是：

$$P_s \sim N\left(p, \frac{pq}{n}\right)$$



知道Ps的概率分布很有用处——这表明，在已知特定总体的情况下，我们可以计算样本的成功比例的概率。我们可以用正态分布近似该分布，样本越大，近似结果越正确。

抽样分布的连续性修正

在用正态分布进行上述近似计算时，进行连续性修正十分重要，这是因为样本中的成功数目是离散的，进行比例计算时用到了这个离散值。

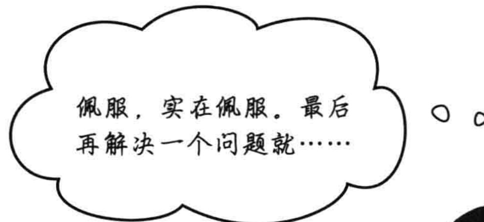
如果用X代表样本中的成功数目，则$P_s=X/n$。X的连续性修正为 ± (1/2)，即连续性修正的算式为：

$$连续性修正 = \frac{\pm 1}{2n}$$

也就是说，如果用正态分布近似计算抽样比例的概率，一定要用 ± 1/2n 进行连续性修正。

有多少糖球？

利用比例的抽样分布，你成功地求出了某一个特定样本中出现一定比例的成功事件的概率。这就是说，现在你可以用样本预测总体情况，或是用所了解的总体信息预测样本情况。





又来了一个问题……

曼帝糖果公司还有一个问题需要你动手解决——除了大盒装糖球，曼帝糖果也生产小袋装糖球，你可以把小袋糖球装在口袋里随身带着，想吃就吃。

根据曼帝糖果公司对总体的统计，每一个小包装袋里的糖球数目均值为10，方差为1。麻烦来了：他们遭到了投诉。一位最忠实的顾客买了30袋糖球，结果发现每袋糖球中的糖球平均数目只有8.5。

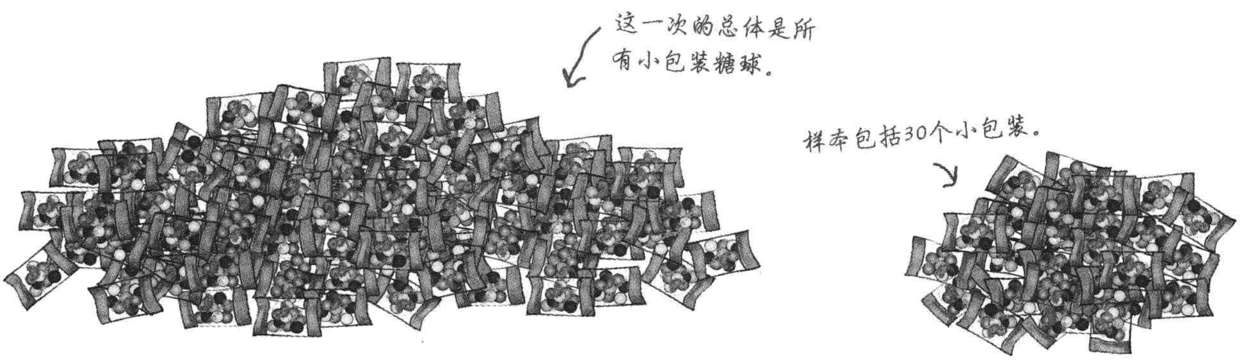
首席执行官担心失去最佳顾客，于是想给他一些补偿，问题是他并不想补偿所有顾客。他想知道，这种事的发生概率有多大？

动动脑

为了解决这类问题，你需要知道什么数据？

我们需要求样本均值的概率

这个问题与前面的问题略有不同。我们已知小包装糖球的总体均值和方差，然后抽取了几袋糖球作为样本，需要为该样本计算概率。这一次，我们不需要计算样本比例的概率，而要计算样本均值的概率。



为了计算样本均值的概率，先要得出样本均值的概率分布。下面是具体步骤：

1. **查看与我们所研究的样本大小相同的所有可能样本。**
   如果我们手头的样本大小为n，则需要考虑大小为n的所有可能样本。小包装糖球有30袋，因此这里的n为30。

2. **查看所有样本形成的分布，求出样本均值的期望和方差。**
   每一个样本都各有特点，每个包装袋中的糖球数目有变化。

3. **得知样本均值的分布后，用该分布求出概率。**
   只要知道所有可能样本的均值的分布情况，就能利用该分布求出一个随机样本的均值的概率，在本例中，随机样本即小包装糖球。

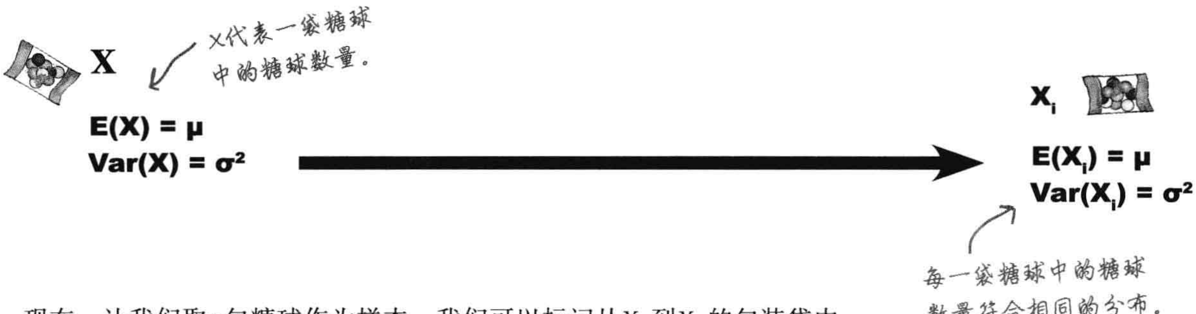
让我们看看如何解决以上问题。

均值的抽样分布

我们如何求样本均值的分布？

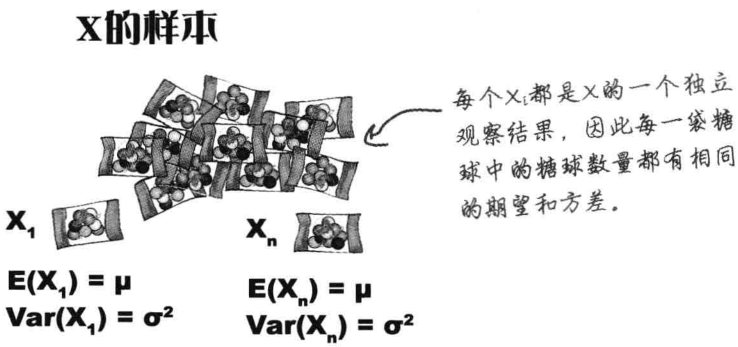
让我们从袋装糖球的总体开始。我们已知总体的均值和方差，并用μ和σ^2表示，一个包装袋中的糖球数量可以用X表示。

随机选择的每一袋糖球都是X的一个**独立观察结果**，因此，每一袋糖球都符合相同的分布。即，如果用X_i代表随机选择的一袋糖球中的糖球数量，则每个X_i的期望都是μ，方差都是σ^2。



现在，让我们取n包糖球作为样本，我们可以标记从X_1到X_n的包装袋中的糖球数量，每个X_i都是X的一个独立观察结果，于是它们遵守相同的分布；每一个X_i的期望都是μ，方差都是σ^2。

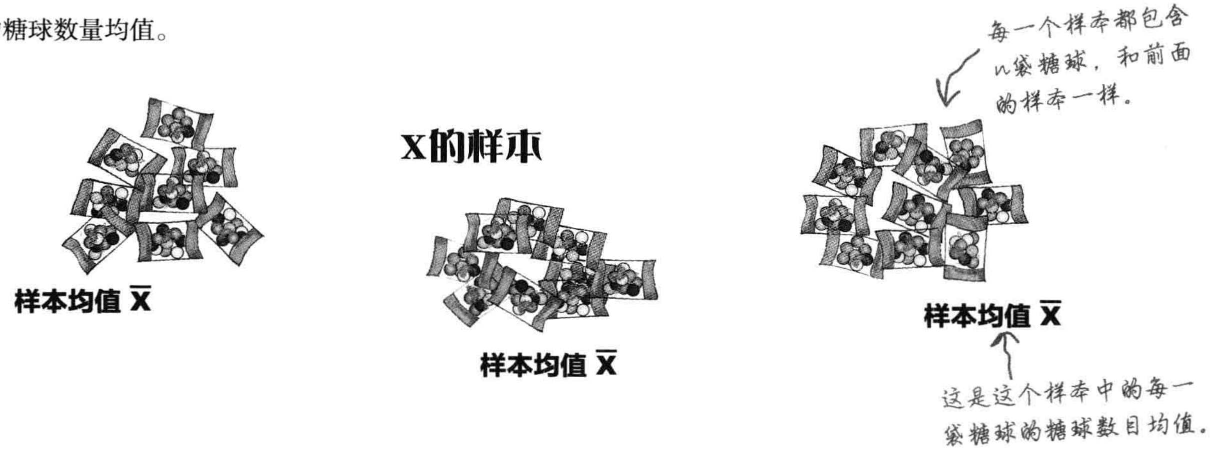
我们可以用$\overline{X}$表示这n袋糖球的容量均值，$\overline{X}$的数值取决于n袋糖球中的每一袋糖球的容量，计算时，要将所有糖球的数量加起来，再除以n。



这是样本均值，是各个包装袋中的糖球容量的平均数。

$$\overline{X} = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

可以取出的大小为n的可能样本为数众多。每一个可能样本都包含n袋糖球，即每一个样本都包含X的n个独立观察结果；每一个随机选择的包装中的糖球数量都遵守相同的正态分布；我们以相同的方法计算每一个样本的糖球数量均值。



我们可以利用从所有可能样本得出的所有样本均值形成一个分布，叫做“均值的抽样分布”，或叫做$\overline{X}$的分布。



均值的抽样分布为我们提供了一种计算样本均值的概率的方法。

为了计算任何一个变量的概率，先要知道这个变量的概率分布，所以，若要计算样本均值的概率，就需要知道样本均值的分布。我们的例子是这样的：在一个有30袋糖球的样本中，求糖球数目的均值小于或等于8.5的概率。

和比例的抽样分布一样，为了能够动手计算概率，先要知道分布的期望和方差。

求$\overline{X}$的期望

前面讲过如何构建均值的抽样分布，即考虑所有大小为n的可能样本，然后用这些样本的均值形成一个分布。

为了能用分布求出概率，先要求出$\overline{X}$的期望和方差。让我们先求$E(\overline{X})$。

这里的$\overline{X}$是样本中的每一袋糖球的容量均值，即：

$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

其中X_i代表第i袋糖球的容量，我们可以利用它求出$E(\overline{X})$。

$$E(\overline{X}) = E\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$

这两个表达式是一样的，只是写法变了变。

$$= E\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right)$$

$$= E\left(\frac{1}{n}X_1\right) + E\left(\frac{1}{n}X_2\right) + \cdots + E\left(\frac{1}{n}X_n\right)$$

可以将这个式子拆分为n个单独的期望，因为：$E(X + Y) = E(X) + E(Y)$.

每一个期望都包含1/n，因此可以从表达式中提取出来，依据为$E(aX) = aE(X)$.

$$= \frac{1}{n}\left(E(X_1) + E(X_2) + \ldots + E(X_n)\right)$$

即，只要我们知道每一个X_i的期望，就能得出$E(\overline{X})$的表达式。

这里的每一个X_i都是X的一个独立观察值，且我们已知$E(X) = \mu$，也就是说，可以用μ代替上式中的各个$E(X_i)$。

于是得到什么呢？

让我们用μ代替各个$E(X_i)$。

$$E(\overline{X}) = \frac{1}{n}(\mu + \mu + \ldots + \mu)$$

X的期望是μ，$E(X_i) = \mu$ 适用于每一个i。

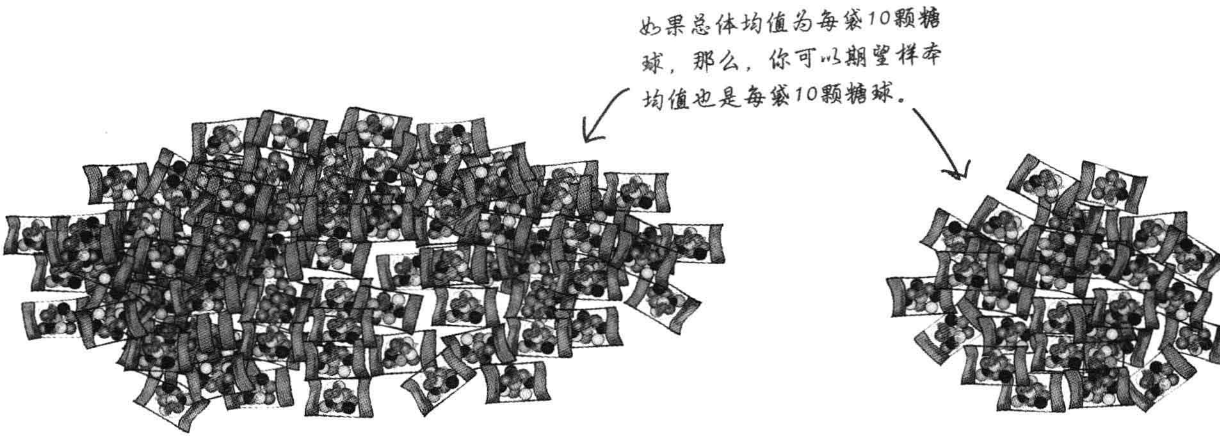
$$= \frac{1}{n}(n\mu)$$

有n个

$$= \mu$$

也就是说E(X)=μ，即所有大小为n的可能样本的均值的平均数等于作为样本来源的总体的均值——实际上，你所求的是所有可能均值的均值。

其实这十分符合直觉——总的看来，你会期望一个样本的每袋糖球平均容量等于总体的每袋糖球平均容量。在我们的具体例子中，总体的每袋糖球平均容量为10，因此，我们会期望样本也是如此。



动动脑

为了求出样本均值的概率，我们还需要知道什么？你认为该怎么求？

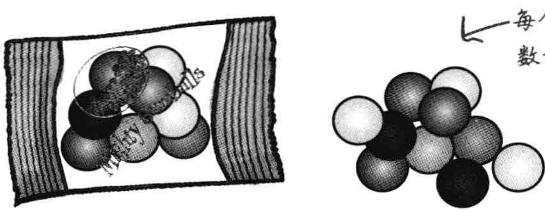
$\overline{X}$的方差是多少？

前面得出了$E(\overline{X})$的算法，不过，在计算样本均值的概率之前，我们还需要求出$Var(\overline{X})$，这样就能朝着$\overline{X}$的分布再迈近一步。

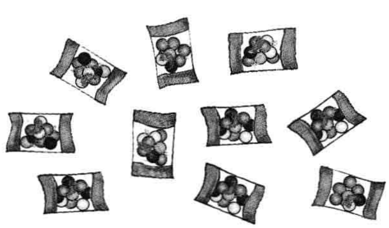


$\overline{X}$的分布不同于X的分布。

X代表一个包装袋中的糖球数量，我们已知一个包装袋中的糖球数目均值，且已知方差。



$\overline{X}$代表一个样本的糖球容量均值，因此$\overline{X}$的分布代表所有可能样本的均值的分布。$E(\overline{X})$表示所有样本均值的均值，而$Var(\overline{X})$指的是样本均值的变异情况。



求$Var(\overline{X})$的过程其实与求$E(\overline{X})$的过程十分类似。



统计量磁贴

通过下面这些算式可求出样本均值的方差的表达式。可惜，有一部分算式掉落了。你的任务是将磁贴放回原位，然后推导出样本均值的方差。

提示：回头复习$E(\overline{X})$的计算过程，这可能会对你有所帮助。

$$Var(\overline{X}) = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$

$$= Var\left(__________\right)$$

$$= Var\left(______\right) + Var\left(______\right) + \cdots + Var\left(______\right)$$

$$= ________ \ (Var(X_1) + Var(X_2) + \cdots + Var(X_n))$$

$$= \frac{1}{n^2}(________)$$

$$= n \times \frac{1}{n^2}\sigma^2$$

$$= ________$$

$\sigma^2 + \sigma^2 + \cdots + \sigma^2$

$\frac{\sigma^2}{n}$

$\frac{1}{n}X_1$

$\frac{1}{n}X_2$

$\frac{1}{n}X_n$

$\left(\frac{1}{n}\right)^2$

$\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n$



统计量磁贴

通过下面这些算式可求出样本均值的方差的表达式。可惜，有一部分算式掉落了。你的任务是将磁贴放回原位，然后推导出样本均值的方差。

$$\mathrm{Var}(\overline{X}) = \mathrm{Var}\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$

$$= \mathrm{Var}\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right)$$

$$= \mathrm{Var}\left(\frac{1}{n}X_1\right) + \mathrm{Var}\left(\frac{1}{n}X_2\right) + \ldots + \mathrm{Var}\left(\frac{1}{n}X_n\right)$$

$$= \left(\frac{1}{n}\right)^2 (\mathrm{Var}(X_1) + \mathrm{Var}(X_2) + \cdots + \mathrm{Var}(X_n))$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2)$$

$$= n \times \frac{1}{n^2}\sigma^2$$

$$= \frac{\sigma^2}{n}$$

做到这一步已经很好。推导过程确实有些曲折，不过我们已经求出了 $\overline{X}$ 的方差——我们知道样本均值会有多大差异。



放轻松

要是完不成这个练习也别灰心，这个练习十分难。

大多数考试委员会都不会要求推导这个算式，你只要记住结果就行了，我们只是为了让你看看这个算式的来历。

均值的抽样分布细细看

让我们好好看看均值的抽样分布。

先看总体X的分布，X的均值为μ，方差为σ^2，因此$E(X) = \mu$而$Var(X) = \sigma^2$。

接着用来自总体X的所有大小为n的可能样本，形成所有样本均值的分布——$\overline{X}$的分布。这个分布的均值和方差计算如下：

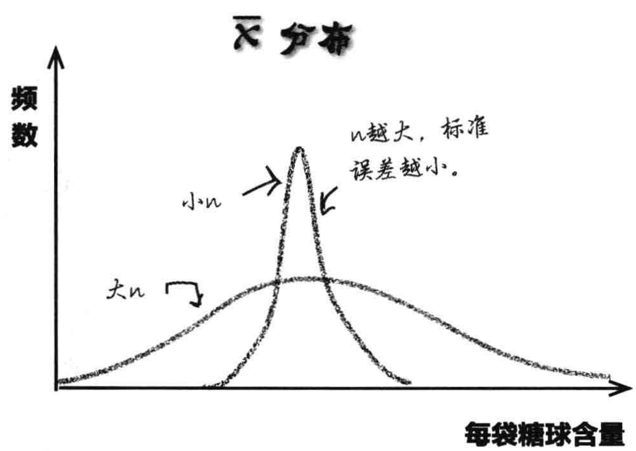
$$E(\overline{X}) = \mu$$

$$Var(\overline{X}) = \frac{\sigma^2}{n}$$

$\overline{X}$的标准差即方差的平方根，这个标准差可指出样本均值与μ的可能偏离距离，因此被称为**均值标准误差**。

$$\text{均值标准误差} = \frac{\sigma}{\sqrt{n}}$$

n越大，均值标准误差越小。也就是说，样本中的个体越多，作为总体均值的估计量的样本均值越可靠。

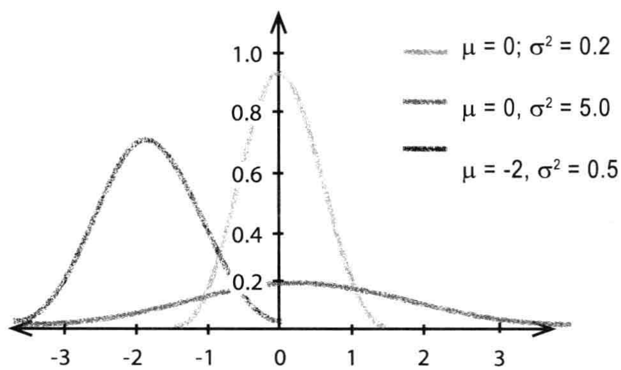


$\overline{X}$如何分布？

前面我们求出了的方差和期望，但还要知道$\overline{X}$的分布，才能求出概率。

让我们先看X为正态分布时，$\overline{X}$符合哪种分布。

下面是各种μ、σ^2及n所对应的分布图，其中X符合正态分布。你注意到什么了？



各种情况下的$\overline{X}$均符合正态分布，也就是说：

这是我们前面求出的$\overline{X}$的均值和方差。

如果 $X \sim N(\mu, \sigma^2)$，则 $\overline{X} \sim N(\mu, \sigma^2/n)$



X可能不符合正态分布。

为了算出样本均值的概率，我们需要知道$\overline{X}$的分布情况，问题是，我们并不知道X如何分布。

我们需要知道，如果X不符合正态分布，$\overline{X}$符合什么分布。

当n很大时，$\overline{X}$仍然可以用正态分布近似

随着n增大，$\overline{X}$越来越接近正态分布。我们已经知道，当X符合正态分布时，$\overline{X}$符合正态分布；如果X不符合正态分布，但如果n足够大，我们仍然可以用正态分布近似的分布。

现在的情况是，我们知道总体的均值和方差，但却不知道总体的分布。不过，这没关系，由于样本大小为30,我们还是能用正态分布求$\overline{X}$的概率。

这叫做“中心极限定理”。

认识中心极限定理

中心极限定理是指：如果从一个非正态总体X中取出一个样本，且样本很大，则$\overline{X}$的分布近似为正态分布。如果总体的均值和方差为μ和σ^2，且n很大，例如大于30，则：

$$\overline{X} \sim N(\mu, \sigma^2/n)$$

这是$\overline{X}$的均值和方差。

是不是觉得很熟悉？这和X符合正态分布时的情况是一样的。唯一的差别是，当X符合正态分布时，样本的大小无所谓。

根据中心极限定理，如果X的样本很大，则$\overline{X}$的分布近似为正态分布。

使用中心极限定理

在实践中，中心极限定理有什么作用呢？让我们看一看。

二项分布

假设你有一个总体，用$X \sim B(n, p)$表示，其中n大于30。如前所述，$\mu = np$，$\sigma^2 = npq$.

根据中心极限定理，在这种情况下，$\overline{X} \sim N(\mu, \sigma^2/n)$。为了求出$\overline{X}$的分布，我们代入总体的数值，即，代入$\mu = np$和$\sigma^2 = npq$，得到：

$$\overline{X} \sim N(np, pq)$$

对于二项分布，总体均值为np，方差为npq，如果将这些式子代入抽样分布，则得到$\overline{X} \sim N(np, pq)$。

泊松分布

现在，假设总体符合泊松分布$X \sim Po(\lambda)$，n还是大于30。对于泊松分布来说，$\mu = \sigma^2 = \lambda$。

和以前一样，我们可以借助正态分布求出σ^2的概率。如果将以上总体参数代入$\overline{X} \sim N(\mu, \sigma^2/n)$，得到：

$$\overline{X} \sim N(\lambda, \lambda/n)$$

对于泊松分布来说，均值和方差都为λ，将这些参数代入抽样分布，得到$\overline{X} \sim N(\lambda, \lambda/n)$。

一般情况下，会使用分布$\overline{X} \sim N(\mu, \sigma^2/n)$，并代入$\mu$和$\sigma^2$的数值。

求出概率

由于$\overline{X}$符合正态分布，于是可以用标准正态概率表查找概率，也就是说，其他正态分布的算法完全适用于你的情况。

练习

让我们用以上结论解决曼帝糖果公司的问题。

每袋糖球的均值为10，方差为1，如果抽取一个有30袋糖球的样本，那么样本均值小于等于8.5(颗/袋)的概率是多少？请按照给出的步骤进行计算。

1. $\overline{X}$符合哪种分布?

2. $P(\overline{X} < 8.5)$的数值是多少?



让我们用以上结论解决曼帝糖果公司的问题。

每袋糖球的均值为10，方差为1，如果抽取一个有30袋糖球的样本，那么样本均值小于等于8.5(颗/袋)的概率是多少？请按照给出的步骤进行计算。

1. $\overline{X}$符合哪种分布？

我们已知 $\overline{X} \sim N(\mu, \sigma^2/n)$，$\mu = 10$，$\sigma^2 = 1$，$n = 30$，而$1/30 = 0.0333$。于是得到：

$\overline{X} \sim N(10, 0.0333)$

2. $P(\overline{X} < 8.5)$的数值是多少？

由于$\overline{X} \sim N(10, 0.0333)$，我们需要求8.5的标准分，以便能够在概率表中查找结果。得到：

$$z = \frac{8.5 - 10}{\sqrt{0.0333}}$$

$$= -8.22$$ (保留两位小数)

$P(Z < z) = P(Z < -8.22)$

这个概率太小了，因此未出现在概率表中。我们可以认为概率如此之小的事件几乎不会出现。

世上没有傻问题

问： 中心极限定理要求进行任何连续性修正吗?

答： 问得好，回答是：不用。你使用中心极限定理求出的概率与样本均值有关，而与样本中的数值无关。因此不需要进行任何连续性修正。

问： 点估计量和抽样分布之间有关系吗?

答： 有关系。让我们先看均值。总体均值的点估计量为$\overline{X}$，即$\hat{\mu}=\overline{X}$。那么均值的抽样分布的期望则为$E(\overline{X})=\mu$。全部样本均值的期望等于μ，我们可以用样本均值估计μ。

与此相似，总体比例的点估计量为P_s，即样本比例，也就是说$p=P_s$。如果我们取全部样本比例的期望，可得$E(P_s)=p$。全部样本比例的期望等于p，于是我们可以用样本比例估计p。

对于方差，我们就不打算在这里进行证明了，但结果相似，即：

$\sigma^2 = s^2$，$E(S^2) = \sigma^2$。

问： 这是巧合吗?

答： 这并非巧合，估计量是这样选择的：以同样方法抽取大小为n的大量样本，使得这些样本的期望等于总体参数的真值。如果做到了这一点，我们就说这些估计量是无偏估计量。

无偏估计量有可能准确无误，这是因为，从所有可能样本的平均情况上看，可以期望该估计量等于真实的总体参数。

问： 标准误差与此有何关系?

答： 总体参数的最佳无偏估计量通常为方差最小的估计量，即标准误差最小的估计量。

要点

- 如果考虑同一个总体中所有大小为n的可能样本，然后用这些样本的均值形成分布，则该分布为“均值的抽样分布”，我们用$\overline{X}$表示样本均值随机变量。
- $\overline{X}$的期望和方差的定义式为：

  $$E(\overline{X}) = \mu$$
  $$Var(\overline{X}) = \sigma^2/n$$

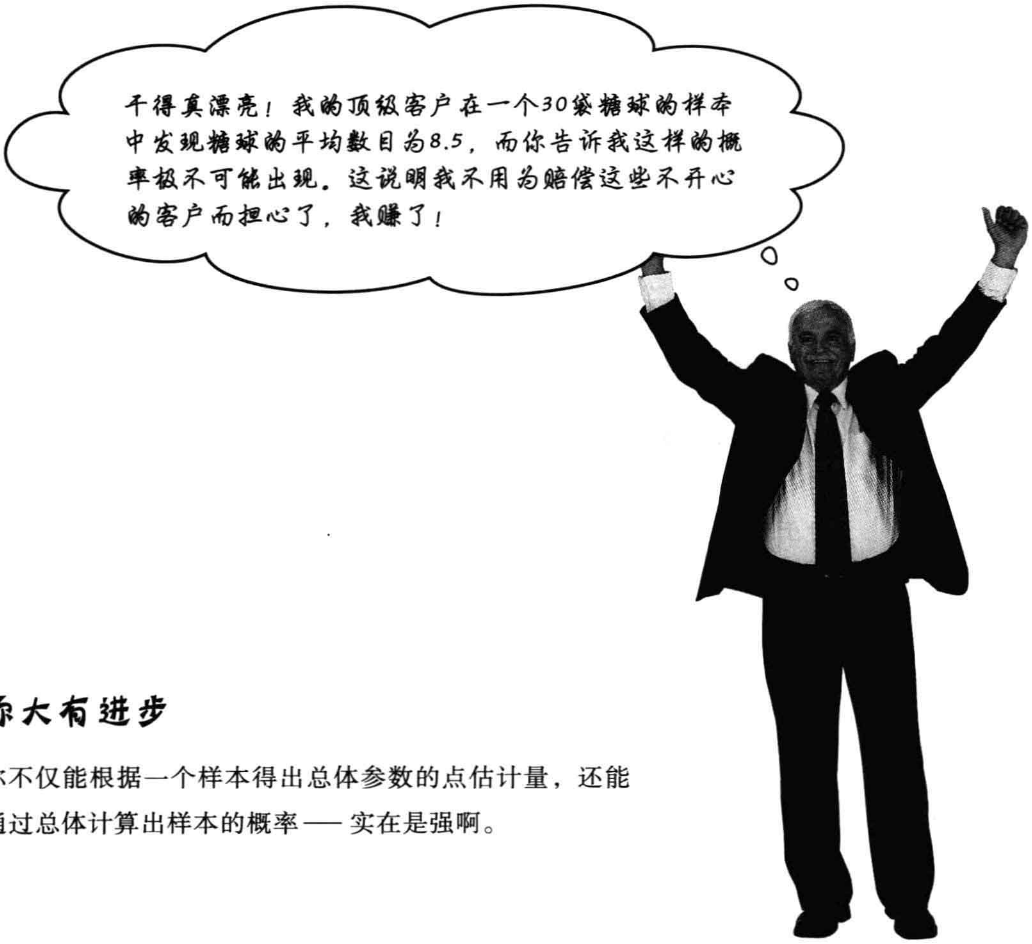
  其中μ和σ^2为总体的均值和方差。
- “均值的标准误差”等于该分布的标准差，即：

  $$\sqrt{Var(\overline{X})}$$

  如果$X \sim N(\mu, \sigma^2)$，则$\overline{X} \sim N(\mu, \sigma^2/n)$.
- 中心极限定理说的是：如果n很大且X不符合正态分布，则：

  $$\overline{X} \sim N(\mu, \sigma^2/n)$$

抽样结果扭转乾坤！



你大有进步

你不仅能根据一个样本得出总体参数的点估计量，还能通过总体计算出样本的概率——实在是强啊。

12 置信区间的构建

自信地猜测



有时候样本无法给出足够正确的结果。

前面讲到如何用点估计量估计总体均值、方差或一定比例的精确值。问题在于，你怎么能肯定自己的估计完全正确？毕竟，你仅仅依靠一个样本对总体作出假设，如果这个样本出问题怎么办？本章将介绍**另一种估计总体统计量的方法** —— 一种**考虑了不确定性的方法**。拿出你的概率表，我们将向你讲解**置信区间**的来龙去脉。

曼帝糖果出事了

曼帝糖果公司的首席执行官大做广告，他言之凿凿、满怀骄傲地宣布了超长效糖球的口味持续时间——精确到秒。

可是……

我们碰到麻烦了。有人自行作了测试，得出了不同的结果。他们威胁说要告我们，这可是要花钱的。



曼帝糖果公司用一个包含100粒糖球的样本得出口味持续时间均值的点估计量为62.7分钟，同时总体方差的点估计量为25分钟。首席执行官在电视节目黄金时段宣布：糖球口味的平均持续时间为62.7分钟。这是根据手头证据有可能得出的最可靠的口味持续时间估计，可要是略有差池，那该怎么办？

如果有人因为曼帝糖果公司的广告和他们打官司，公司就会又赔钱又丢生意。他们需要你帮忙摆脱困境。

他们需要你出手相救。

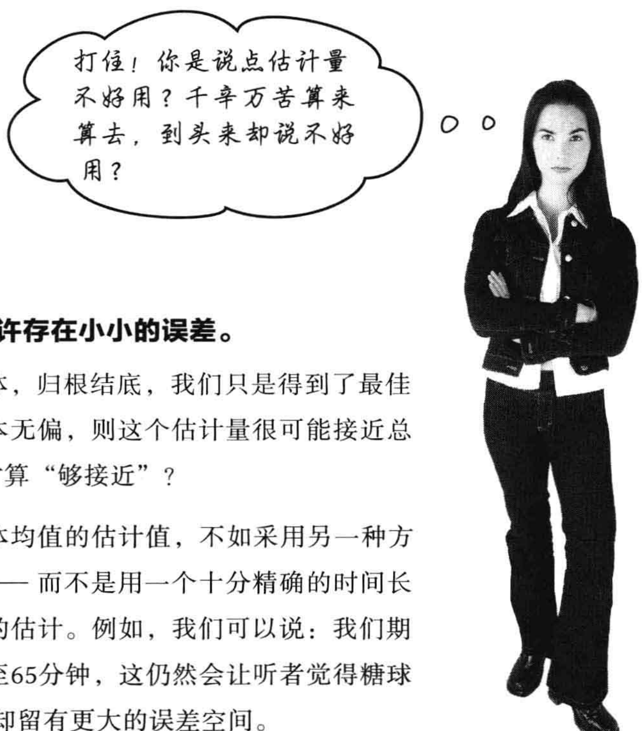
动动脑

你认为错在哪里？曼帝糖果公司是否应该用点估计量的精确值做广告？为什么？

精度引起的问题

如上一章所讲，点估计量是我们有可能给出的总体统计量的最佳估计。你取用最具代表性的数据样本，以此估计总体的主要统计量，如均值、方差、比例，这意味着超长效口香糖球的口味持续时间均值的点估计量是我们有可能给出的最佳估计。

点估计量的推导过程存在这样的问题：我们依赖来自唯一的一个样本的结果得出非常精确的估计。我们想了很多办法，确保样本无偏，使样本尽量具有代表性；但对于这个样本是不是能100%地代表总体，我们并没有绝对的把握，原因很简单——我们用的是样本。



点估计量是有价值的，但也许存在小小的误差。

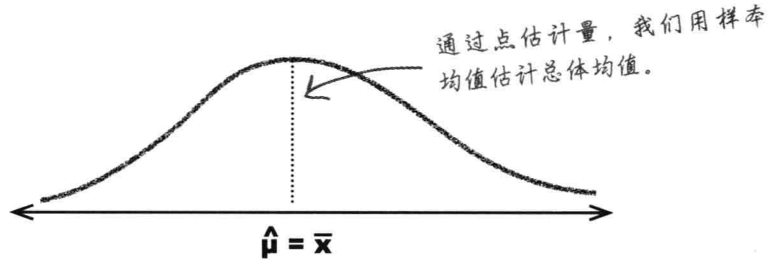
由于我们并没有使用整个总体，归根结底，我们只是得到了最佳估计量。如果我们所用的样本无偏，则这个估计量很可能接近总体的真值。问题是，多接近才算“够接近”？

与其给出一个精确值作为总体均值的估计值，不如采用另一种方法。我们可以指定某个区间——而不是用一个十分精确的时间长度，作为糖球口味持续时间的估计。例如，我们可以说：我们期望糖球的口味持续时间为55至65分钟，这仍然会让听者觉得糖球口味持续时间接近1小时，但却留有更大的误差空间。

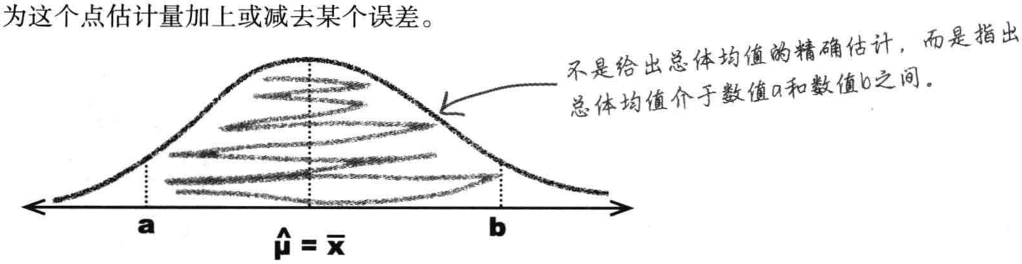
问题是，我们如何确定区间？这就看你希望自己对结果有多大自信了……

认识置信区间

此前，我们以样本数据为基础，利用点估计量估计了糖球口味持续时间的均值，通过点估计量，我们能够给出糖球口味平均持续时间的非常精确的估计。下面这张图体现了糖球样本口味持续时间的分布。



那么，如果我们为总体均值指定一个区间，情况会怎么样呢？我们不指定一个确切的数值，而指定两个数值——我们期望糖球口味持续时间介于这两个数值之间。我们让均值的点估计量处于这个区间的中央，并将这个区间的上下限设定为这个点估计量加上或减去某个误差。



选择区间上下限是为了让“总体均值介于a和b之间”这一结果具有特定概率。例如，你可能希望通过选择a和b，使得该区间中包含总体均值的几率为95%。也就是说，所选择的a和b使得：

$$P(a < \mu < b) = 0.95$$

我们用(a,b)表示这个区间，由于a和b的确切数值取决于你希望自己对于“该区间包含总体均值”这一结果具有的可信程度，因此，(a, b)被称为**置信区间**。

那么，我们如何求总体均值的置信区间？

求解置信区间四步骤

下面是求解置信区间的几大步骤。要是没办法一下子弄明白每个步骤的目的，别担心，我们很快会具体讲解。

1. **选择总体统计量** ← 是指希望用于构建置信区间的总体统计量。
2. **求出其抽样分布** ← 上一章讲过抽样分布。
3. **决定置信水平** ← 你选择的区间中包含该统计量的概率
4. **求出置信上下限** ← 为了求出置信上下限，我们需要知道置信水平和抽样分布。

让我们看看是否能够替曼帝糖果首席执行官构建一个可以进行广告宣传的置信区间——让我们求出糖球口味持续时间均值的置信区间。

世上没有傻问题

问： 你能为任何一个总体统计量构建一个置信区间吗？

答： 一般说来，只要知道抽样分布，就能为任何总体统计量构建置信区间。我们已经讲过均值和比例的抽样分布，因此能够为这两个统计量构建置信区间。

问： 方差呢？我们能为方差构建置信区间吗？

答： 理论上是可以的，不过我们还没有讲过方差的分布，也不打算讲。较为常见的做法是构建均值和比例的置信区间，统计学考试往往考这些内容。

问： 上面这些步骤是和均值的置信区间有关系还是和比例的置信区间有关系？

答： 这些步骤对于二者是通用的——既可以用于总体均值，也可以用于总体比例。

问： 总体的分布情况是否有关系？

答： 关键在于你要为之构建置信区间的统计量的抽样分布，如果想求均值的置信区间，就要知道均值的抽样分布；如果想求比例的置信区间，就要知道比例的抽样分布。

总体分布对置信区间的主要影响在于它对抽样分布的影响。我们随后加以阐述。

第1步：选择总体统计量

第1步是选取要为之构建置信区间的统计量，这取决于要解决的实际问题。

在我们的实例中，需要为口香糖球口味持续时间的均值构建一个置信区间，于是就需要为总体均值μ构建一个置信区间。

选好总体统计量，就可以进行下一步了。

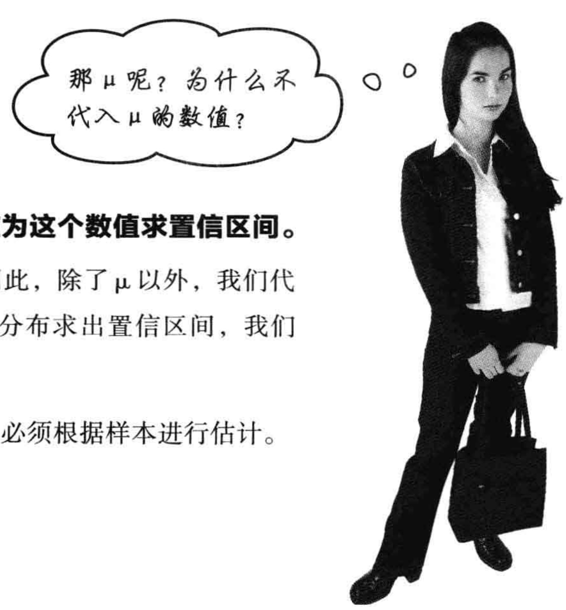
第2步：求出所选统计量的抽样分布

为了求出总体均值的抽样分布，我们需要知道均值的抽样分布，即需要知道$\overline{X}$的期望和方差以及其分布。

让我们先求期望和方差。回顾上一章的内容，我们知道均值的抽样分布的期望和方差为：

$$E(\overline{X}) = \mu \qquad Var(\overline{X}) = \frac{\sigma^2}{n}$$

为了利用以上结果求出μ的置信区间，我们代入总体方差的数值σ^2和样本大小的数值n。



我们不代入μ的数值，这是因为我们正在为这个数值求置信区间。

我们正在利用抽样分布求μ的置信区间，因此，除了μ以外，我们代入所有数值。代入σ^2和n之后，就能用$\overline{X}$的分布求出置信区间，我们很快就会进行说明。

只有一个问题——我们并不知道σ^2的真值，必须根据样本进行估计。

点估计量出手相救

那么用哪个数值作为 σ^2 值呢？

尽管我们不知道总体方差 σ^2 的真实值，却可以用它的点估计量进行估计。于是我们代入 $\hat{\sigma}^2$，或者叫做 s^2，而不是 σ^2。

于是均值的抽样分布的均值和方差等于：

$$E(\overline{X}) = \mu \qquad Var(\overline{X}) = \frac{s^2}{n}$$

这是方差的点估计量。我们不知道总体方差的真实值是多少，于是用样本方差进行估计。

曼帝糖果公司用包含100颗糖球的样本计算估计值，并算得 $s^2 = 25$，于是：

$$Var(\overline{X}) = \frac{s^2}{n} = \frac{25}{100} = 0.25$$

还有一事待定：为了能求出 μ 的置信区间，我们需要清楚地知道 $\overline{X}$ 的分布。

动动笔

假定 $X \sim N(\mu, \sigma^2)$，且样本包含的数目很大。$\overline{X}$ 符合哪种分布？用前面算出的 $E(\overline{X})$ 和 $Var(\overline{X})$ 来帮忙。

动动笔解答

假定$X \sim N(\mu, \sigma^2)$，且样本包含的数量很大。$\overline{X}$符合哪种分布？用前面算出的 $E(\overline{X})$ 和 $Var(\overline{X})$ 来帮忙。

如果X符合正态分布，那么$\overline{X}$也符合正态分布，代入σ^2的点估计量，得到：

$$\overline{X} \sim N(\mu, s^2/n)$$

或

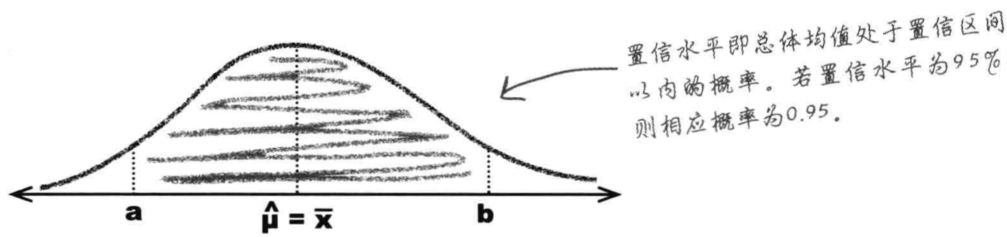
$$\overline{X} \sim N(\mu, 0.25)$$

我们已经求出了$\overline{X}$的分布

既然已经知道 $\overline{X}$ 的分布情况，我们就有了足够的信息，可以进入下一步。

第3步：决定置信水平

置信水平表明你希望自己对于“置信区间包含总体统计量”这一说法有多大把握。例如，假设我们希望总体均值的置信水平为95%，这表示总体均值处于置信区间中的概率为0.95。



动动脑

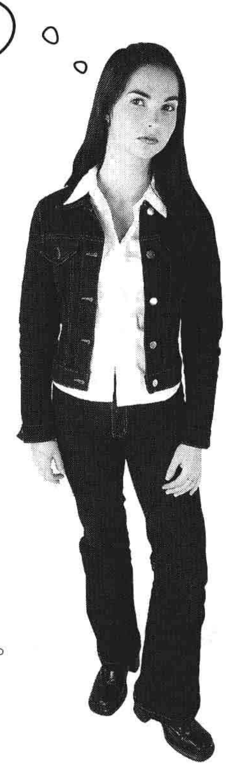
你觉得置信水平对置信区间的大小有何影响？

如何选择合适的置信水平

那么由谁来决定置信水平？多大的置信水平才合适？

答案完全取决于你的具体情况以及你需要对“区间中包含总体统计量”这一说法有多大信心。常用的置信水平是95%，但有时候你可能会另有要求，如90%或99%。例如，曼帝糖果公司首席执行官希望对“总体均值位于置信区间之中”这一说法有更大的信心，这样他才能在电视中广而告之。

关键是记住这一点：置信水平越高，区间越宽，置信区间包含总体统计量的几率越大。



把置信区间弄得太宽的问题是：置信区间会失去其意义。

举个极端例子：我们可以说糖球口味持续时间的均值在0至3天之间。这固然不错，但你却无法据此知道糖球口味实际上能持续多久 —— 不知道持续时间是几秒，几分钟，还是几小时。

关键在于，要让区间尽可能窄，但又要足够宽，这样才能合理地相信真正的均值就在区间之中。

让我们为曼帝糖果公司选用95%的置信水平，如此一来，区间中包含总体均值的概率就很高了。

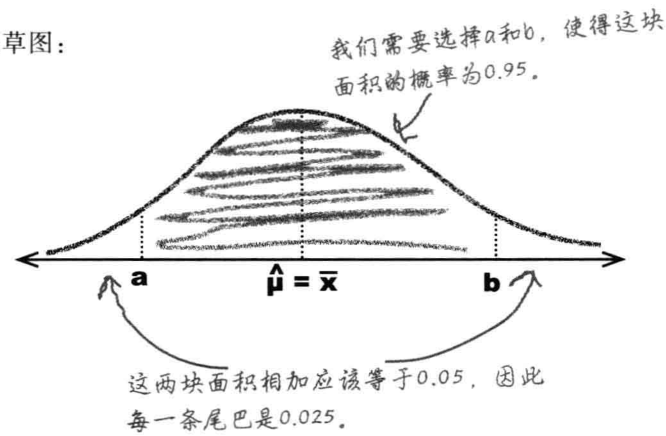
既然已经选定了置信水平，就可以进入最后一步：求出置信上下限。

第4步：求出置信上下限

最后一步是求a和b —— 置信区间的**上下限**，上下限指出一个范围的左右边界 —— 均值有95%的概率落入这个范围中。a和b的确切值取决于需要使用的抽样分布以及需要具有的置信水平。

对于我们的实例，需要让糖球口味持续时间均值具有95%的置信度，即，μ 位于我们求得的a和b之间的概率必须为0.95。我们还知道，$\overline{X}$ 符合正态分布，其中 $\overline{X} \sim N(\mu, 0.25)$。

下面是我们需要使用的一张草图：



利用 $\overline{X}$ 的分布我们可以求出a和b的值。即，我们可以利用 $\overline{X} \sim N(\mu, 0.25)$ 求出a和b，例如 $P(\overline{X} < a) = 0.025$ 和 $P(\overline{X} > b) = 0.025$。



由于 $\overline{X}$ 符合正态分布，所以我们可以用正态分布求置信区间。

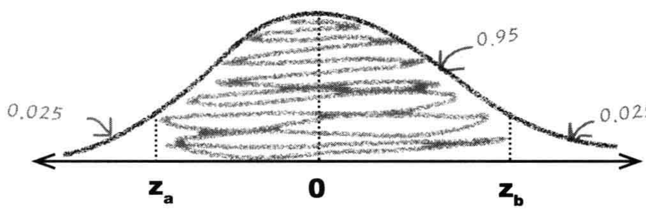
具体算法和前面讲过的其他问题的算法相似：算出标准分，查询标准正态分布概率表，得出所需要的结果。

先求Z

为了能够利用正态分布表，先对$\overline{X}$进行标准化。我们已知 $\overline{X} \sim N(\mu, 0.25)$，于是，经过标准化计算，得到：

$$Z = \frac{\overline{X} - \mu}{\sqrt{0.25}} \qquad \text{其中} \qquad Z \sim N(0, 1)$$

下面是经过标准化的置信区间图形：



我们需要求出 z_a 和 z_b，其中 $P(z_a < Z < z_b) = 0.95$，即标准置信上下限为 z_a 和 z_b，其中 $P(Z < z_a) = 0.025$ 且 $P(Z > z_b) = 0.0255$。利用概率表可以求出 z_a 和 z_b 的值。

动动笔

我们需要求出 z_a 和 z_b，使得 $P(z_a < Z < z_b) = 0.95$。

1. 使用概率表求出Z_a的数值，使得$P(Z < z_a) = 0.025$。

2. 使用概率表求出Z_b的数值，使得$P(Z > z_b) = 0.025$。

动动笔解答

我们需要求出 z_a 和 z_b，使得 $P(z_a < Z < z_b) = 0.95$。

1. 使用概率表求出Z_a的数值，使得$P(Z < z_a) = 0.025$。

   在标准概率表中查找0.025，得$z_a = -1.96$。

2. 使用概率表求出Z_b的数值，使得$P(Z > z_b) = 0.025$。

   对于Z_b，需要查找0.975，得$z_b = 1.96$。

用μ改写不等式

到此为止，我们求出了置信区间的标准形式，得到

$P(-1.96 < Z < 1.96) = 0.95$，即：

$$P\left(-1.96 < \frac{\overline{X} - \mu}{0.5} < 1.96\right) = 0.95$$



用μ改写不等式，即可以得到μ的置信区间。

如果将

$$-1.96 < \frac{\overline{X} - \mu}{0.5} < 1.96$$

改写为这种形式：

$$a < \mu < b$$

这个式子给出了μ的区间。

就能得到μ的上下限。



奇妙池

你的任务是改写 $-1.96<(\overline{X}-\mu)/0.5<1.96$，得出 μ 的置信区间。从池中取出零星公式，放在空白的横线上。每一个公式碎片的使用次数**不得**超过一次。

这是不等式左边。→ $-1.96 < \frac{\overline{X}-\mu}{0.5} < 1.96$ ← 这是不等式右边。

左边：

$-1.96 < \frac{\overline{X}-\mu}{\ldots\ldots}$

$-1.96 \times \ldots\ldots < \overline{X}-\mu$

$\ldots\ldots + \mu < \overline{X}$

$\mu < \ldots\ldots\ldots$

右边：

$\frac{\overline{X}-\mu}{\ldots\ldots} < 1.96$

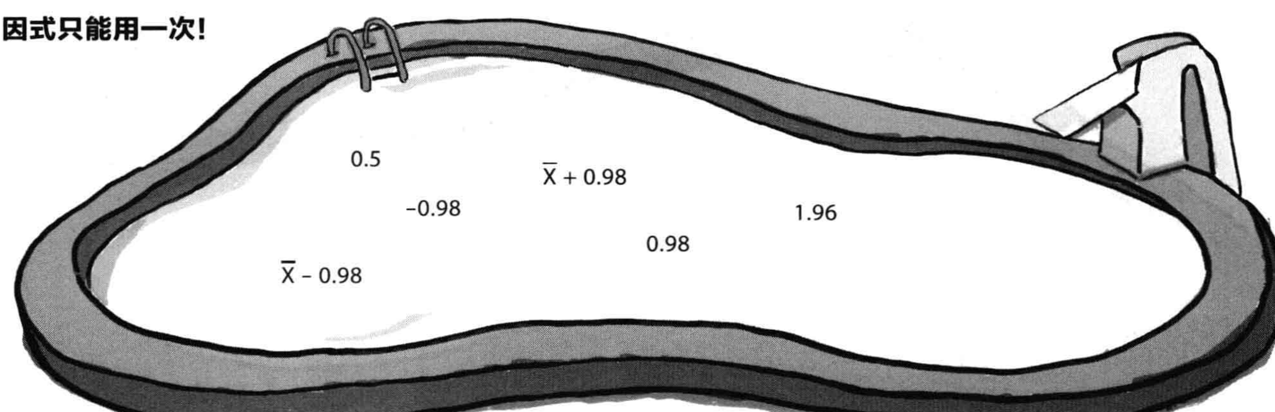
$\overline{X}-\mu < \ldots\ldots \times 0.5$

$\overline{X} < \ldots\ldots + \mu$

$\ldots\ldots\ldots < \mu$

$\overline{X}-0.98 < \mu < \overline{X}+0.98$ ← 这是综合起来的结果。

说明：池中的每一个因式只能用一次！



奇妙池解答



你的任务是改写$-1.96<(\overline{X}-\mu)/0.5< 1.96$，得出 μ 的置信区间。从池中取出零星公式，放在空白的横线上。每一个公式碎片的使用次数不得超过一次。

这是不等式左边。

$$-1.96 < \frac{\overline{X}-\mu}{0.5} < 1.96$$

这是不等式右边。

$$-1.96 < \frac{\overline{X}-\mu}{0.5}$$

$$-1.96 \times 0.5 < \overline{X}-\mu$$

$$-0.98 + \mu < \overline{X}$$

$$\mu < \overline{X} + 0.98$$

$$\frac{\overline{X}-\mu}{0.5} < 1.96$$

$$\overline{X}-\mu < 1.96 \times 0.5$$

$$\overline{X} < 0.98 + \mu$$

$$\overline{X} - 0.98 < \mu$$

$$\overline{X} - 0.98 < \mu < \overline{X} + 0.98$$

说明：池中的每一个因式只能用一次！

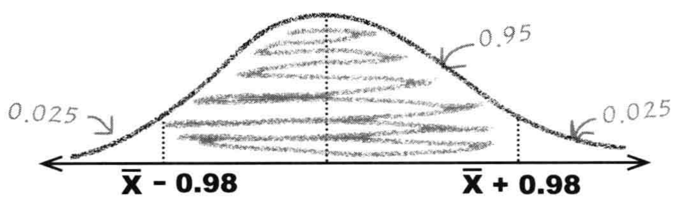


最后求 $\overline{X}$ 的数值

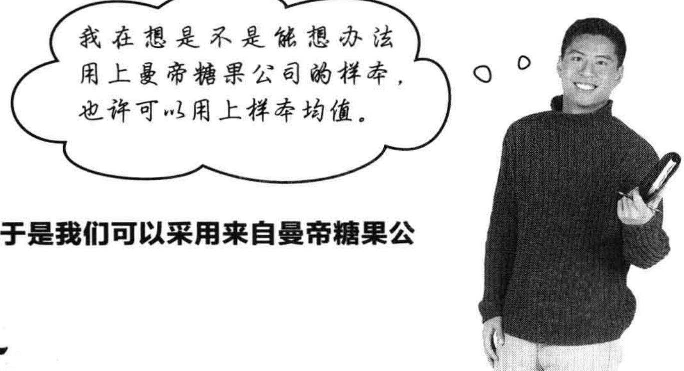
写出不等式后，我们就非常接近描述糖球典型口味持续时间的数值——μ的置信区间。即，我们使用：

$$P(\overline{X} - 0.98 < \mu < \overline{X} + 0.98) = 0.95$$

下面是草图：



我们的置信上下限为 $\overline{X} - 0.98$ 和 $\overline{X} + 0.98$，只要知道用哪个数值作为 $\overline{X}$，就能得出置信上下限。



$\overline{X}$ 指的是样本均值的分布，于是我们可以采用来自曼帝糖果公司样本的 $\overline{x}$ 值。

动动笔

置信上下限分别为 $\overline{X} - 0.98$ 和 $\overline{X} + 0.98$，对于曼帝糖果公司的样本，$\overline{x}$ 为62.7。请使用这个数值求出置信上下限的数值。

动动笔解答

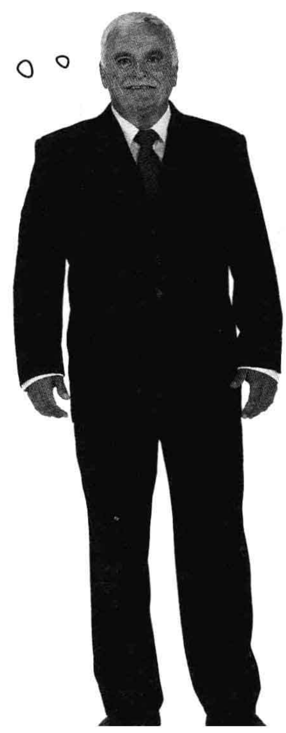
置信上下限分别为 $\overline{X} - 0.98$ 和 $\overline{X} + 0.98$，对于曼帝糖果公司的样本，$\overline{x}$ 为62.7。请使用这个数值求出置信上下限的数值。

置信上下限分别为 $\overline{X} - 0.98$ 和 $\overline{X} + 0.98$，如果代入样本均值，则置信上下限等于62.7 − 0.98 和62.7 + 0.98，即置信区间为(61.72, 63.68)。

你求出了置信区间

祝贺！你旗开得胜，求出了一个置信区间。你的结论是：区间(61.72,63.68)中包含糖球口味持续时间总体均值的几率是95%。

超级棒的消息！这么说我能更新那些漂亮的广告用语了，这就不存在打官司的问题了。



首席执行官在电视广告中用置信区间取代了点估计量，给出了对糖球口味持续时间的准确而精确的估计，却不必提到精确的数字——就算样本有误差也还有周旋余地。

步骤总结

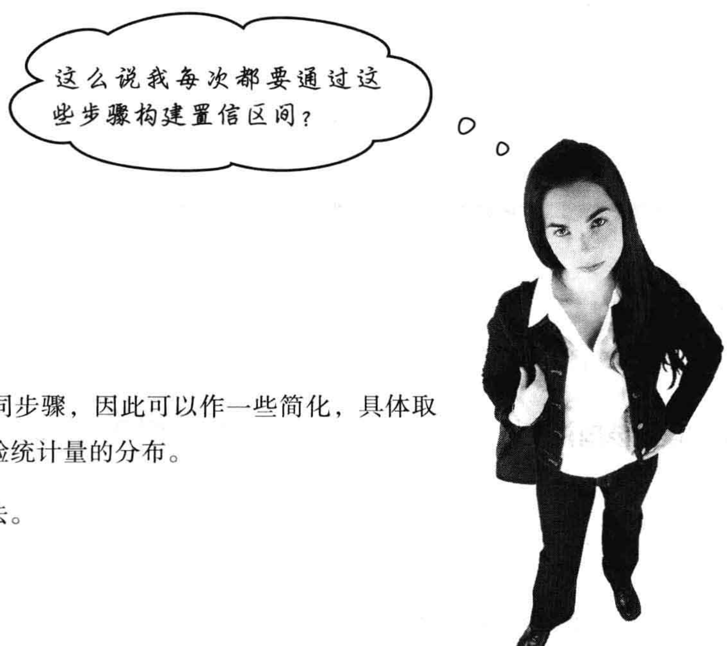
让我们复习一下前面讲过的置信区间的构建步骤。

首先**选择用于构建置信区间的总体统计量**。我们需要求出糖球口味持续时间均值的置信区间，于是需要构建 μ 的置信区间。

确定了用于构建置信区间的总体统计量后，接着**求其抽样分布**。我们求得均值的抽样分布的期望和方差，代入除 μ 以外的各个统计量的数值，于是发现我们可以使用 $\overline{X}$ 的正态分布。

随后，我们确定了用于构建置信区间的置信水平 —— 95%。

最后必须**求出置信区间的置信上下限**。我们利用置信水平和抽样分布得出了合适的置信区间。



我们可以作一些简化。

构建置信区间会反复使用相同步骤，因此可以作一些简化，具体取决于所需要的置信水平和试验统计量的分布。

让我们看看其中一些简化方法。

置信区间简便算法

下面是一些实用的置信区间简便算法。你只要查看要求的总体统计量、总体分布以及各种条件，然后代入总体统计量或其估计量，就行了。数值c取决于置信水平。

总体统计量	总体分布	条件	置信区间
μ	正态	σ^2已知 n可大可小 $\bar{x}$为样本均值	$\left(\bar{x}-c\frac{\sigma}{\sqrt{n}}, \bar{x}+c\frac{\sigma}{\sqrt{n}}\right)$
μ	非正态	σ^2已知 n很大（至少30） $\bar{x}$为样本均值	$\left(\bar{x}-c\frac{\sigma}{\sqrt{n}}, \bar{x}+c\frac{\sigma}{\sqrt{n}}\right)$
μ	正态或非正态	σ^2未知 n很大（至少30） $\bar{x}$为样本均值 s^2为样本方差	$\left(\bar{x}-c\frac{s}{\sqrt{n}}, \bar{x}+c\frac{s}{\sqrt{n}}\right)$
p	二项	n很大 p_s为样本比例 q_s等于$1-p_s$	$\left(p_s-c\sqrt{\frac{p_sq_s}{n}}, p_s+c\sqrt{\frac{p_sq_s}{n}}\right)$

c的数值取决于所需要的置信水平。只要以正态分布作为试验基础，就可以用这些数值。

置信水平	C值
90%	1.64
95%	1.96
99%	2.58

一般如何计算区间？

一般情况下，置信区间的计算式为：

统计量±(误差范围)

误差范围等于c与检验统计量的标准差的乘积。

误差范围 = c × (统计量的标准差)

练习

曼帝糖果公司抽取了一个大小为50的样本，发现样本中的红色糖球的比例为0.25。请为总体中具有这一比例的红色糖球构建一个置信水平为99%的置信区间。



曼帝糖果公司抽取了一个大小为50的样本，发现样本中的红色糖球的比例为0.25。请为总体中具有这一比例的红色糖球构建一个置信水平为99%的置信区间。

总体比例的置信区间为：

$$\left(p_s - c\sqrt{\frac{p_s q_s}{n}},\ p_s + c\sqrt{\frac{p_s q_s}{n}}\right)$$

我们需要求99%置信水平的置信区间，因此c=2.58。红色糖球的比例为0.25，于是p_s=0.25且q_s=0.75，n=50，于是得出：

$$\left(p_s - c\sqrt{\frac{p_s q_s}{n}},\ p_s + c\sqrt{\frac{p_s q_s}{n}}\right) = \left(0.25 - 2.58\sqrt{\frac{0.25 \times 0.75}{50}},\ 0.25 + 2.58\sqrt{\frac{0.25 \times 0.75}{50}}\right)$$

$$= (0.25 - 2.58 \times 0.0612,\ 0.25 + 2.58 \times 0.0612)$$

$$= (0.25 - 0.158,\ 0.25 + 0.158)$$

$$= (0.092,\ 0.408)$$

世上没有傻问题

问： 之前求$\overline{X}$的期望和方差的时候，为什么代入σ^2的点估计量，却不代入μ的点估计量？

答： 由于我们需要求的正是μ的置信区间，因此不用$\overline{x}$代替μ。我们需要求出含有μ的表达式，以便求出置信区间。

问： 为什么用$\overline{x}$作为$\overline{X}$的值？

答： $\overline{X}$的分布即均值的抽样分布。它是这样来的：从总体中取出每一个大小为n的可能样本，然后用所有的样本均值形成一个抽样分布。

$\overline{x}$是来自样本的特定均值，于是我们借助它求出置信区间。

问： 置信区间和置信水平有何区别？

答： 置信水平是“统计量处于置信区间之中”的概率，通常是一个百分数，例如95%。置信区间则给出了区间本身——数字实际范围的上下限。

问： 我们已经求得μ的95%置信区间为(61.72, 63.68)，这究竟意味着什么？

答： 这意味着：如果你打算抽取大小相同的多个样本，然后为所有这些样本构建置信区间，则这些置信区间中有95%会包含总体均值的真实值。你由此知道，用这种方法构建的置信区间在95%的情况下都将包含总体均值。

问： 简便算法中的c适用于所有置信区间吗？

答： 它们适用于所有我们讲过的简便算法，这是因为这些简便算法都基于正态分布——所给出的各种条件下的抽样分布都符合正态分布。

问： 我曾经看到置信区间的简便算法中用的是“a”而不是“c”，有错吗？

答： 完全没错。关键在于，无论你把这个数字叫做“a”还是叫做“c”，它所代表的总是你代入置信区间以便达到合适的置信水平的那个数——无论如何称呼，数字总是一样的。

问： 是否所有的置信区间都基于正态分布？

答： 并非如此。我们随后会讲到基于其他分布的区间。

问： 既然只要在简便算法中代入数值就行，为什么讲那么多步骤呢？

答： 讲这些步骤是为了让你看清楚问题实质，理解置信区间的构建过程。大多数时候，你只要代入数值就行了。

问： 使用置信区间时需要进行连续性修正吗？

答： 理论上是要的，不过实践中常忽略不计，也就是说只要在简便算法中代入数值算出置信区间就行了。



还有一个问题……

曼帝糖果公司最后还有一个问题需要你解决。有一家糖果店想知道糖球的典型重量，原因是他们发现顾客往往按照重量购买糖球，而不是按照数量购买。要是糖果店知道糖球的典型重量，就能利用这个信息进行促销。



曼帝糖果公司抽取了一个具有代表性的样本，共10颗，然后称了每一粒糖球的重量。这个样本的 $\bar{x} = 0.5$ 盎司，$s^2 = 0.09$。

我们如何求出置信区间？

第1步：选择总体统计量

第1步是选取要为其构建置信区间的统计量。我们需要为糖球重量均值构建一个置信区间，也就是要为总体均值 μ 构建置信区间。

由于需要求 μ 的置信区间，于是下一步就是求 μ 的抽样分布 —— $\overline{X}$ 的分布。

动动脑

假设总体中的每一粒糖球的重量都符合正态分布，你如何为这个数据建立一个95%置信区间？提示：查看前面的置信区间简便算法一览表，看看我们符合哪种条件。

第2步：求 $\overline{X}$ 的概率分布

那么，$\overline{X}$符合什么分布呢？



并非任何情况都能用正态分布进行良好近似。

我们前面讲过的所有抽样分布要么符合正态分布，要么可以用正态分布进行近似。问题是，我们无法对每一个置信区间都使用正态分布。不巧，目前碰上的就是这种不能用正态分布的情况。

不能用正态分布的原因何在？

当抽样很大时，正态分布是求解置信区间的理想分布 —— 能得出精确结果，且与总体本身是否是正态分布无关。

现在我们碰到了另一种情况 —— 尽管X本身符合正态分布，$\overline{X}$却并不符合。



主要原因有二。

第一，我们不知道总体方差的确切值，因此必须利用样本数据估计σ^2，我们可以通过点估计量轻松地完成这项工作，但是，还有第二个原因：样本太小，估计值很有可能出现较大误差 —— 比使用大样本的误差要大得多。这些潜在的误差意味着使用正态分布无法得出足够精确的 $\overline{X}$ 的概率，那样就无法得出精确的置信区间。

那么，$\overline{X}$符合哪种分布呢？实际上，它符合t分布。让我们具体看看。

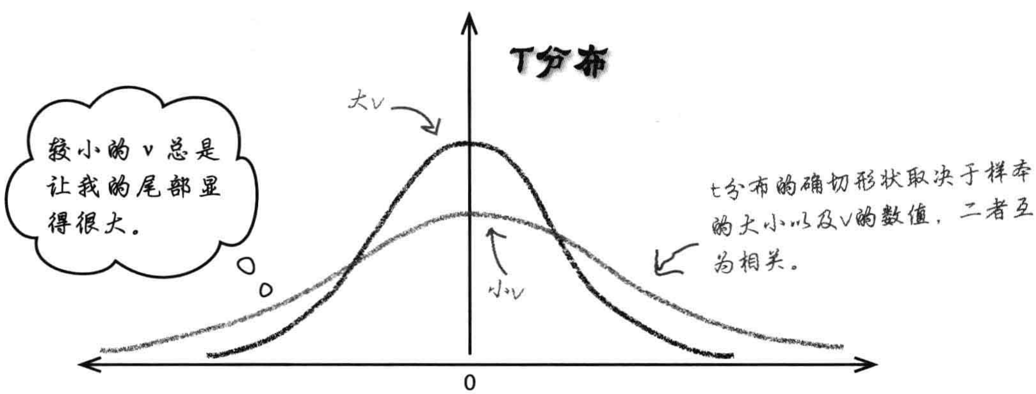
当样本很小时，$\overline{X}$符合t分布

当总体符合正态分布，σ^2未知，且可供支配的样本很小时，$\overline{X}$符合t分布——这种分布正好可以用来处理我们面临的问题。

t分布是外形光滑、对称的曲线，确切形状取决于样本大小。当样本很大时，t分布外形很像正态分布；当样本很小时，曲线较为扁平，有两条粗粗的尾巴。它只有一个参数——ν，$\nu = n-1$。n为样本的大小，ν被称为**自由度**。

我们会在第14章中更深入地探讨自由度。

让我们看看下面这张图：这是各种ν对应的t分布。你能看出ν对分布形状有什么影响吗？



“T符合t分布且自由度为v”的简明表示方法为：

$$T \sim t(v)$$

T为检验统计量，计算方法见下一页。

t(v)表示：我们正在使用自由度为v的t分布；v = n - 1。

t分布的使用方法与正态分布相似——先将概率区间的上下限转化为标准分，然后用概率表求出所需要的结果。

让我们先求标准分。

求t分布的标准分

t分布的标准分的计算方法与正态分布的标准分的计算方法相同。像处理正态分布一样，我们先减去抽样分布的期望，然后用所得到的差除以标准差。唯一的差别是，我们用T而不是Z代表结果，这是为了配合t分布的使用。

我们需要求出 $\overline{X}$ 的分布，于是要用到 $\overline{X}$ 的期望和标准差。$\overline{X}$ 的期望为μ，标准差为σ/n。由于需要用s估计σ的数值，于是t分布的标准分的算式如下：

这个公式和Z的计算公式一样——减去均值，除以标准差。→

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

这是总体均值，我们正在求其置信区间。

这是 $\overline{X}$ 的标准差。

我们只要代入 $\overline{X}$、$\hat{\sigma}$ 和 n 就行了。

动动笔

让我们看看如何将以上结果应用于曼帝糖果的抽样：
抽样中共有10粒糖球，其中$\overline{x}$ = 0.5盎司，s^2 = 0.09。
ν 的数值是多少？T值又是多少？

动动笔解答

让我们看看如何将以上结果应用于曼帝糖果的抽样：
抽样中共有10粒糖球，其中$\bar{x}$ = 0.5盎司，s^2 = 0.09。
ν 的数值是多少？T值又是多少？

样本共有10粒糖球，$\nu = n - 1$，即 ν 的数值为9。

T计算如下：

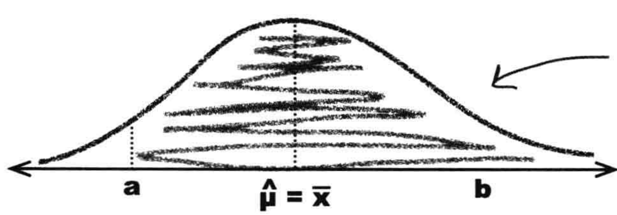
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

$$= \frac{\bar{X} - \mu}{\sqrt{0.09/10}}$$

$$= \frac{\bar{X} - \mu}{0.0949}$$

第3步：决定置信水平

那么该为曼帝糖果选用哪个置信水平呢？记住：置信水平指的是你希望自己对“置信区间包含总体统计量”这个说法有多大信心，它帮助我们指出置信区间应该有多宽。像以前一样，让我们用95%作为总体均值的置信水平，于是总体均值位于置信区间之中的概率为0.95。



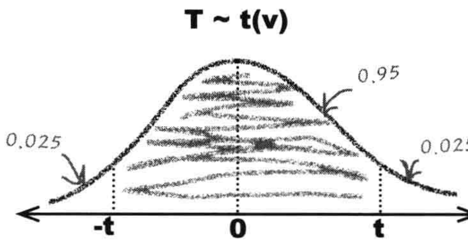
既然已经有了置信水平，我们就能进入最后一步——求 μ 的置信区间。

第4步：求出置信上下限

t分布的置信上下限的算法类似于正态分布的算法，即可通过下式进行计算：

$$\left(\bar{x} - t\frac{s}{\sqrt{n}}, \bar{x} + t\frac{s}{\sqrt{n}}\right)$$

这个式子和前面见过的式子是一样的，只不过用t代替了c。



其中

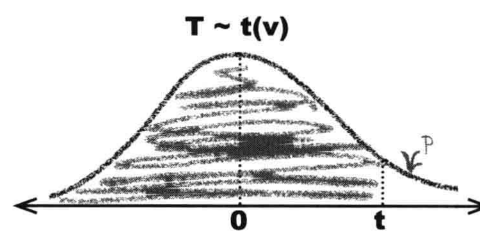
$$P(-t \le T \le t) = 0.95$$

等于0.95，这是因为我们希望求95%置信区间。

我们可以通过t分布概率表求出t值。

使用t分布概率表

通过t分布概率表可求出P(T > t) = p中的t值。在我们的实例中，p = 0.025。



为了求出t值，先从概率表中查找第一列的ν值，再查找第一行的p值，二者的交点处即为t值。例如，查找ν =7和p=0.05，可得t=1.895。

求出t值后，就能求置信区间了。

p = 0.05

	尾部概率 p											
v	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	[illegible]	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	[illegible]	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	[illegible]	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	[illegible]	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	[illegible]	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	[illegible]	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	[illegible]	[illegible]	[illegible]	[illegible]	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	[illegible]	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	[illegible]	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	[illegible]	2.228	2.359	2.764	3.169	3.581	4.144	4.587

ν = 7

7和0.05相交处。

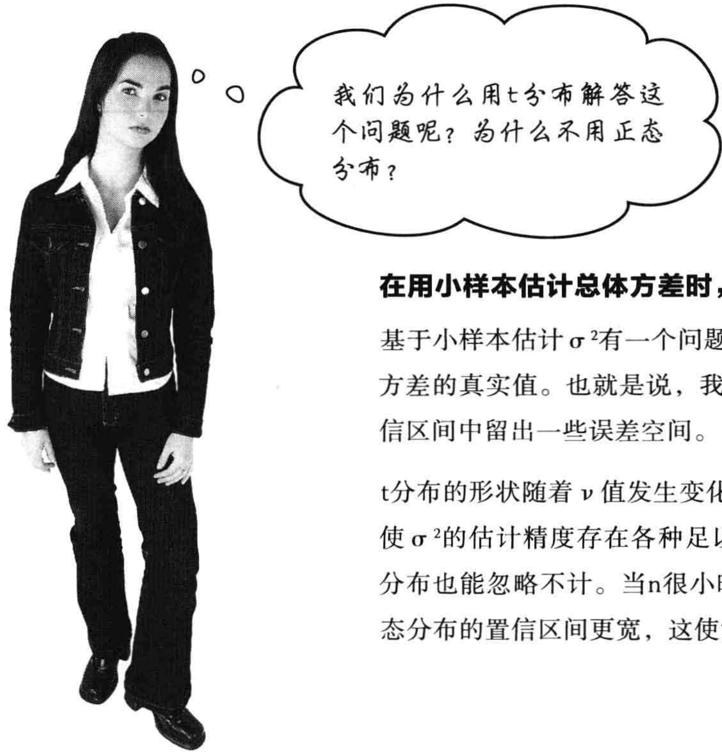
练习

看看能否求出糖球平均重量的95%置信区间。样本包含10粒糖球，且 $\overline{x}$ = 0.5 盎司，s^2 = 0.09。

1. μ 的置信区间等于 $(\overline{x} - t\, s/\sqrt{n}, \overline{x} + t\, s/\sqrt{n})$，使用标准概率表求t值。

2. 用t值求 μ 的置信区间。

t分布与正态分布比较



在用小样本估计总体方差时，t分布更精确。

基于小样本估计σ^2有一个问题，即可能无法精确地反映总体方差的真实值。也就是说，我们需要让区间变宽，以便在置信区间中留出一些误差空间。

t分布的形状随着ν值发生变化，由于考虑了样本的大小，即使σ^2的估计精度存在各种足以让人有所察觉的不确定性，t分布也能忽略不计。当n很小时，t分布给出的置信区间比正态分布的置信区间更宽，这使它更适合用于小样本。

置信区间简明算法－t分布

下面是有关t分布的使用时机以及μ的置信区间的简单提示。

总体统计量	总体分布	条件	置信区间
μ	正态或非正态	σ^2未知 n很小（小于30） $\bar{x}$为样本均值 s^2为样本方差	$\left(\bar{x} - t(v)\frac{s}{\sqrt{n}}, \bar{x} + t(v)\frac{s}{\sqrt{n}}\right)$

为了求出$t(\nu)$，需要查找t分布概率表。为此，用$\nu=n-1$和你确定下来的置信水平求出置信区间。



看看能否求出糖球平均重量的95%置信区间。样本包含10粒糖球，且 $\bar{x}$ = 0.5 盎司，s^2 = 0.09。

1. μ 的置信区间等于 $(\bar{x} - t\, s/\sqrt{n}, \bar{x} + t\, s/\sqrt{n})$，使用标准概率表求t值。

样本中有10粒糖球，因此 ν = 9。我们希望求出95%置信区间，即需要在t分布概率表中查找0.025，自由度为9。于是得出：t = 2.262。

2. 用t值求 μ 的置信区间。

我们将 $\bar{x}$、t、s和n代入 $(\bar{x} - t\, s/\sqrt{n}, \bar{x} + t\, s/\sqrt{n})$ 求置信区间，得到：

$$(\bar{x} - t\, s/\sqrt{n}, \bar{x} + t\, s/\sqrt{n}) = (0.5 - 2.262 \times \sqrt{(0.09/10)}, 0.5 + 2.262 \times \sqrt{(0.09/10)})$$

$$= (0.5 - 2.262 \times 0.0949, 0.5 + 2.262 \times 0.0949)$$

$$= (0.5 - 0.215, 0.5 + 0.215)$$

$$= (0.285, 0.715)$$

练习

曼帝糖果公司发现他们的装糖机出问题了。他们抽取了30台机器作为样本，发现故障次数均值是15。请为每月故障次数构建一个99%置信区间。



曼帝糖果公司发现他们的装糖机出问题了。他们抽取了30台机器作为样本，发现故障次数均值是15。请为每月故障次数构建一个99%置信区间。

每月故障次数符合泊松分布模型，由于有30台机器，我们可以用$(\bar{x} - cs/\sqrt{n}, \bar{x} + cs/\sqrt{n})$求解置信区间。

我们需要求99%置信区间，于是$c = 2.58$。泊松分布的期望和方差都等于λ，因此$\bar{x} = 15$且$s^2 = 15$。

置信区间计算如下：

$$
\begin{aligned}
(\bar{x} - cs/\sqrt{n}, \bar{x} + cs/\sqrt{n}) &= (15 - 2.58 \times \sqrt{(15/30)}, 15 + 2.58 \times \sqrt{(15/30)}) \\
&= (15 - 2.58 \times \sqrt{(15/30)}, 15 + 2.58 \times \sqrt{(15/30)}) \\
&= (15 - 2.58 \times 0.707, 15 + 2.58 \times 0.707) \\
&= (15 - 1.824, 15 + 1.824) \\
&= (13.176, 16.824)
\end{aligned}
$$

世上没有傻问题

问： $\bar{x}$符合t分布吗？

答： 当总体符合正态分布而样本很小时，X符合t分布，这时需要使用样本数据估计总体方差。

问： 一般说来，如果置信水平发生改变，对置信区间会有何影响？

答： 如果置信水平下降，则置信区间变窄；如果置信水平提高，则置信区间变宽。例如，对于一组相同的数据，95%置信区间将比99%置信区间更窄。

问： 如果样本大小n发生改变，对置信区间会有何影响？

答： 如果n减小，则置信区间变宽；如果n增大，则置信区间变窄。

置信区间的表达式为：

统计量±误差范围

其中，误差范围 = c x 统计量的标准差。

统计量的标准差取决于样本的大小——n越大，统计量的标准差越小；这就是说，n越大误差范围越小，n越小误差范围越大。

一般说来，较小的样本形成较宽的置信区间，较大的样本形成较窄的置信区间。

置信区间求出来了！

你再本章进步很大——所以现在你有两种估计总体统计量的方法了。

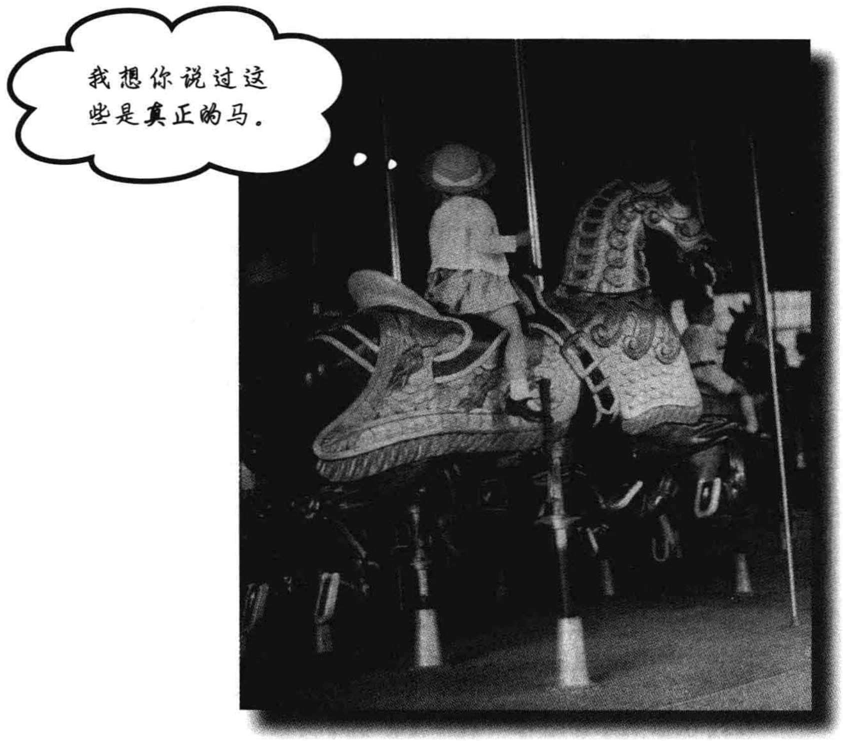
第一种估计方法是使用**点估计量**。点估计量方法可用于估计总体统计量的精确数值，是根据样本数据有可能做出的最好猜测。

另一种估计方法是使用总体统计量的**置信区间**。这个方法得到的并非总体统计量的精确估计，而是求出总体统计量的一个有较高可信度的数值范围。



13 假设检验的运用

研究证据



他人的言论未必句句真实可信。

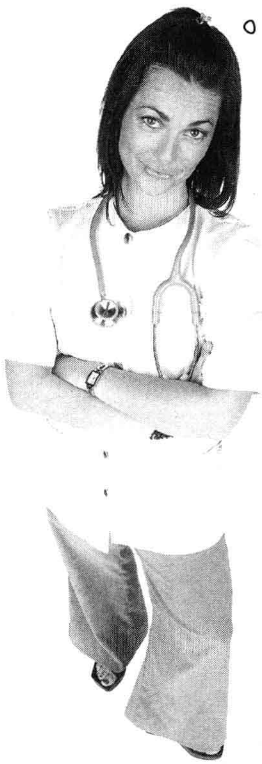
问题是如何判断他人的言论何时真，何时假？假设检验为你提供了一种方法 —— 利用样本检验各种统计断言是否可能属实。通过假设检验可以权衡证据，检验极限结果 —— 是纯属巧合，还是存在其他内在根据？让我们一起阅读本章，看看如何利用假设检验证实或打消你内心深处的疑虑。



统计邦新上市的神奇药品

统计邦头号制药公司生产了一种治疗打鼾的新药物。被打呼噜折磨不堪的患者纷纷赶往医院，指望能得到睡眠救星。

制药公司断言他们的神药能在两周内治愈90%的患者，对于深受打鼾困扰的人来说，这可是个天大的好消息。问题是，并非人人都信服这个断言。



统计邦外科诊所的医生给病人开了鼾克，但她对结果感到失望。她决定自行对药物进行试验。

她随机抽取了15位鼻鼾患者，对这些患者实施为期2周的鼾克疗法。两周后，她请这些患者来医院复诊，看他们是否不再打鼾。

结果如下：

医生只记录了患者的鼻鼾是否治愈。

是否治愈?	是	否
频数	11	4

动动笔

如果药物能治愈90%的鼻鼾患者，那么你会期望这个包含15名鼻鼾患者的抽样中出现几位治愈者？你认为这符合什么分布？

动动笔解答

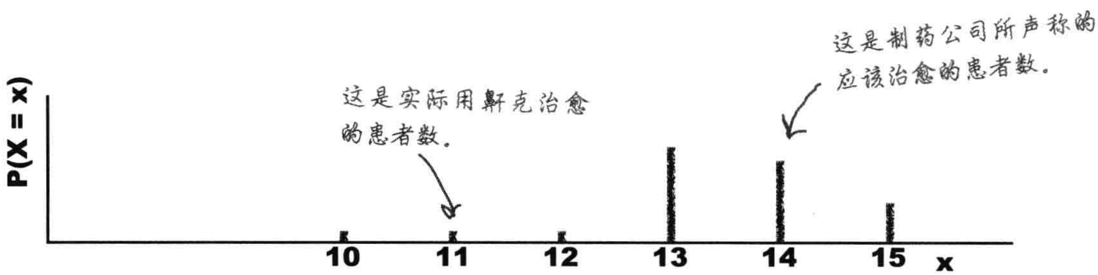
如果药物能治愈90%的鼻鼾患者，那么你会期望这个包含15名鼻鼾患者的抽样中出现几位治愈者？你认为这符合什么分布？

15的90%是13.5，因此你可以期望治愈14名患者。而医生的抽样中只有11名患者治愈，这比期望的结果小得多。

由于试验次数一定，且医生关注的是治愈人数，因此，治愈人数符合二项分布。如果用X表示治愈人数，则 $X \sim B(15, 0.9)$。

问题出在哪里？

下面的概率分布代表制药公司所宣称的能够通过鼻鼾新药治愈的人数。



医生抽样中通过鼾克治愈的患者数实际上比你所期望的治愈人数小得多。按照制药公司的说法，你会期望治愈14名患者，但其实只治愈了11名患者。

为什么会出现这样的差别呢？



制药公司可能不是存心撒谎，但他们的断言可能具有误导性。

制药公司的检验可能有缺陷，进而使得鼾克的断言可能存在误导性——由于疏忽大意，他们对鼾克进行的检验可能有缺陷，或者说有偏差，以致于对总体做出了不准确的预测。

如果鼾克的治愈率实际上低于90%，那么就能解释为什么样本中只有11人治愈。



制药公司的断言实际上可能是准确的。

如果制药公司没有出差错，那么很可能是那位医生的抽样患者无法代表整个鼻鼾患者总体。很有可能鼻鼾药物确实治愈了90%的患者，医生却正好抽中了不治愈人数比例较高的样本。也就是说，医生的样本可能存在某种偏倚，要不然就是因为样本中的患者数目较少。

动动脑

你认为我们该怎么办？我们该相信谁？是相信制药公司的断言，还是相信医生的质疑？

纵观全局

我们该如何裁决医生与制药公司之间的矛盾说法？让我们纵观全局，看看需要做点什么。

我们可以对制药公司的断言进行检验，以期裁决制药公司和医生的矛盾之说。即，我们权且相信制药公司的断言，可是一旦出现强有力的反驳证据，我们就改为站到医生一边。

具体做法：

查看断言

记下制药公司的断言。



查看证据

看看我们需要哪些证据才可以否定制药公司的断言，并把所需的证据和我们手头现有的证据进行比较。方法是：先假设制药公司的断言属实，然后看看医生得到的结果是否有误。



作出决策

根据证据，接受或否定制药公司的断言。



通常以上过程称为**假设检验** —— 做出假设或断言，对照证据进行检验。让我们看看假设检验的一般过程。

假设检验六步骤

下面是假设检验的几个粗略步骤，我们将在后面几页详加说明。

即我们要对其进行试验的断言。

1. **确定要进行检验的假设**
2. **选择检验统计量**

我们需要选取能最有效地对断言进行检验的统计量。

我们需要使用某种确定性水平。

3. **确定用于做决策的拒绝域**
4. **求出检验统计量的p值**

我们需要了解在假定断言为真的情况下，我们的试验结果的可信程度。

5. **查看样本结果是否位于拒绝域内**

接着需要了解试验结果是否位于确定性限值范围中。

6. **作出决策**



我们需要确保对药品断言进行正确的检验，然后才能加以否定。

通过这些步骤，我们明白：在对双方进行公正的裁决，同时将对断言进行公正的试验。我们不想在没有足够证据反驳制药厂断言的情况下拒绝该断言，这说明，需要通过某种方式确定所谓“充分证据”应该包含哪些内容。

第1步：确定假设

你在进行这一步 →

确定要进行检验的假设
选择检验统计量
确定用于做决策的拒绝域
求出检验统计量的p值
查看样本结果是否位于拒绝域内
作出决策

让我们先执行假设检验第1步，了解要进行检验的主要断言，该断言被称为**假设**。

制药公司断言

根据制药公司的断言，鼾克能在2周内治愈90%的患者。除非我们有充分证据进行反驳，否则就要接受这个结论。

我们所检验的这个断言被称为**原假设**，以H_0表示，除非我们有充分证据进行反驳，否则就要接受这个断言。

原假设即你要对其进行检验的断言，除非有足够的证据进行反驳，否则你将接受这个断言。→ H_0

我是原假设，是默认的结论。要是你认为我错了，请给出证据。

鼾克的原假设是什么？

鼾克的原假设即制药公司的断言：鼾克能在两周内治愈90%的患者。除非我们有足够的证据进行反驳，否则应认同这个断言。

我们需要检验鼻鼾药物是否至少能治愈90%的患者，因此原假设为：p = 90%。

这就是鼾克试验的原假设。→ H_0: $p = 0.9$

除非能举出反驳的证据，否则你必须认同我能治愈90%患者的结论。



用什么做备选假设？

前面讲过我们即将检验的断言 —— 原假设，可如果这个假设不为真该怎么办？用什么做备选假设？

医生的见解

医生认为制药公司对疗效的断言过于理想，反而显得不真实 —— 她认为治愈率不会达到90%，低于90%的可能性更大。

与原假设对立的断言被称为**备择假设**，用H_1表示。如果有足够的证据拒绝H_0，我们就接受H_1。

备择假设即在拒绝H_0之后将接受的另一个断言。→ H_1

我是备择假设，如果H_0让你失望，你就得选择被当作“备胎”的我了。

鼾克的备择假设

鼾克的备择假设就是在证实制药公司的断言有假之后要认同的另一断言。如果有足够的证据反驳制药公司的断言，那么有可能医生的断言是对的。

医生认为鼾克治愈的患者少于90%，即备择假设为：$p < 90\%$。

这就是鼾克试验的备择假设。→ H_1: $p < 0.9$

既然我们已经为鼾克的假设检验确定了原假设和备择假设，就可以进行第2步了。

世上没有傻问题

问： 既然我们假设原假设是真实的，为什么后来又要找证据证明它是错误的呢？

答： 进行假设检验实际上是对假设检验的断言进行试验，你对假设保持怀疑，随后，如果有足够的拒绝证据，则进行拒绝。这有点儿像把囚犯带到法官面前接受审判。只有在有足够的证据证明囚犯有罪时，才能进行宣判。

问： 原假设和备择假设必须穷举吗？二者是否应该涵盖所有可能的结果？

答： 不用。例如，我们的原假设是$p=0.9$，备择假设是$p<0.9$，二者都不必考虑$p>0.9$。

问： 这个假设检验的样本是不是太小了？

答： 即使样本很小，我们仍然能够做假设检验，这都取决于你所使用的检验统计量，下面将讲到这个问题。

问： 这么说假设检验就是用来证明断言是否正确的？

答： 假设检验无法给出绝对的证明，你只能在假定原假设为真的前提下，通过假设检验了解观察结果到底有多可靠。如果观察结果极不可能发生，就会成为证明原假设为假的证据。

进行假设检验时，你假定原假设为真；如果有足够的证据反驳原假设，则拒绝原假设，接受备择假设。

第2步：选择检验统计量

你在进行这一步 →

- 确定要进行检验的假设
- **选择检验统计量**
- 确定用于做决策的拒绝域
- 求出检验统计量的p值
- 查看样本结果是否位于拒绝域内
- 作出决策

既然已经完全确定了要进行检验的内容，接着就需要通过某些手段进行检验——这可以借助**检验统计量**实现。

“检验统计量”即用于对假设进行检验的统计量，是与该检验关系最为密切的统计量。

鼾克的检验统计量是哪一个？

我们做假设检验的目的是检验鼾克是否能治愈90%以上的患者。为此，可以根据制药公司的说法查看概率分布，看看抽样中的成功次数是否显著。

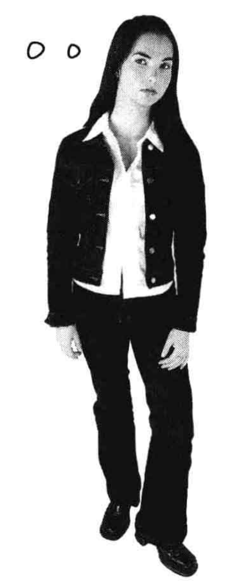
如果用X表示样本人数，就可以将X作为检验统计量。样本中共有15名患者，根据制药公司的说法，成功概率为0.9。由于X符合二项分布，于是检验统计量实际上符合：

我们在524页得出了这个统计量。

$$X \sim B(15, 0.9)$$

这是我们的假设检验的检验统计量。

我糊涂了。为什么我们说成功概率是0.9？我们还不知道是多少呢。



我们根据原假设H_0选择检验统计量。

我们需要检验是否有充足的证据反驳原假设。办法是：首先假设H_0为真，然后寻找不利于H_0的证据。在针对鼾克的检验中，我们假设治愈概率为0.9——除非有有力证据证明这不成立。

为此，我们假定治愈概率为0.9，看看得出观察结果的可能性有多大。也就是说，取样本结果，然后计算发生这个结果的概率——我们通过求**拒绝域**实现这个目标。

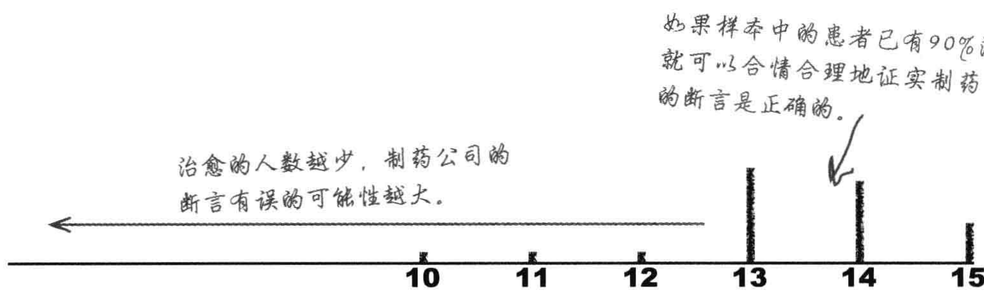
第3步：确定拒绝域

确定要进行检验的假设
选择检验统计量
你在进行这一步 → **确定用于做决策的拒绝域**
求出检验统计量的p值
查看样本结果是否位于拒绝域内
作出决策

假设检验的**拒绝域**是一组数值，这组数值给出反驳原假设的最极端证据。

让我们再看看医生的样本，以便了解拒绝域的使用方法。如果治愈人数为90%或90%以上，这就与制药公司的断言吻合了。随着治愈人数下降，制药公司的断言为真的可能性越来越小。

下面是概率分布：

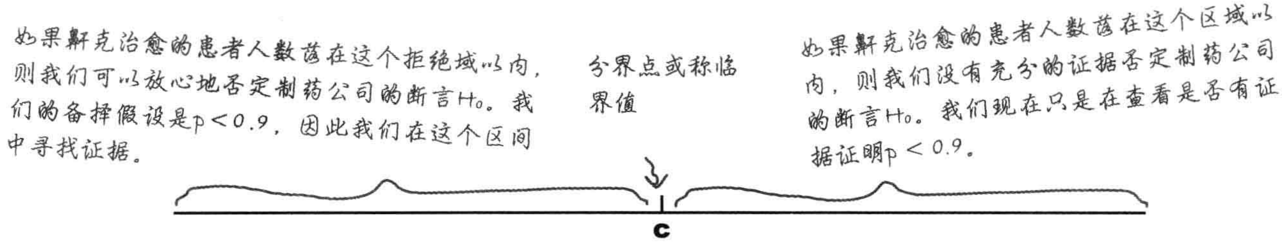


何时能够拒绝制药公司的断言？

样本中得到成功治愈的患者人数越少，可以用于反驳制药公司断言的证据就越有力。问题是：这些证据的强度达到多大时，我们能够坚决地拒绝原假设？—— 到什么程度候能够拒绝“鼾克治愈90%鼻鼾患者”这个断言？

我们需要通过某种方法指出何时能够合理地拒绝原假设 —— 指定一个拒绝域即可实现这一目的。如果鼻鼾患者的治愈人数位于拒绝域以内，我们就说有足够的证据可以反驳原假设；如果鼻鼾患者的治愈人数位于拒绝域以外，我们就承认没有足够的证据可以反驳原假设，并接受制药公司的断言。我们把拒绝域的分界点称为“c”—— **临界值**。

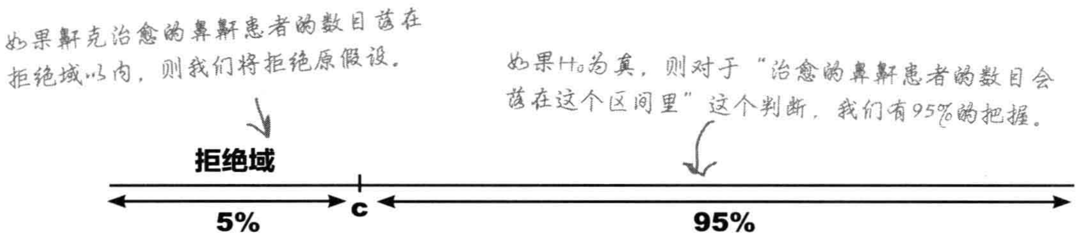
如何选择临界值？



为求拒绝域，先定显著性水平

为了求出假设检验的拒绝域，首先需要定下“显著性水平”。检验的**显著性水平**所量度的是一种愿望，即：希望在样本结果的不可能程度达到多大时，就拒绝原假设H_0。像置信区间的置信水平一样，显著性水平以百分数表示。

例如，假设我们想以5%为显著性水平检验制药公司的断言，这说明我们选取的拒绝域应使得“鼻鼾患者治愈人数小于c”的概率小于0.05，即概率分布最低端的5%部分。



显著性水平通常用希腊字母α表示。α越小，为了拒绝H_0，样本结果需要达到的不可能程度越高。

我们该使用多高的显著性水平？

让我们在假设检验中使用5%的显著性水平。即，如果样本中的治愈患者的数目落在概率分布的最低5%范围内，我们将否定制药公司的断言。如果治愈的鼻鼾患者的数目落在概率分布的95%高端范围内，则我们将判定没有足够的证据反驳原假设，同时接受制药公司的断言。

如果我们用X表示治愈的鼻鼾患者的数目，则我们将拒绝域定义为能令下列不等式成立的一些数值：

$$P(X < c) < \alpha$$

其中

$$\alpha = 5\%$$

重要统计量

显著性水平

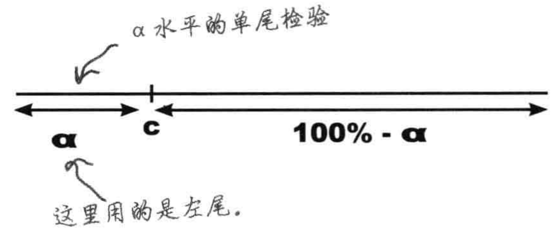
显著性水平用α表示。它表明你希望在观察结果的不可能程度达到多大时拒绝H_0。

拒绝域细细看

在构建检验的拒绝域时，还需要明白一件事：所构建的是**单尾检验**还是**双尾检验**。让我们看看这两者之间的差别，以及它们对拒绝域有什么影响。

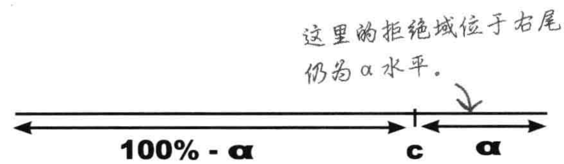
单尾检验

单尾检验即检验的拒绝域落在可能的数据集的一侧，你选择检验水平——以α表示，然后确保拒绝域以相应的概率反映这个水平。尾部可以是可能数据集的左侧或右侧，具体用哪一侧取决于备择假设H_1。



如果备择假设包含一个<符号，则使用**左尾**，此时拒绝域位于数据的低端。

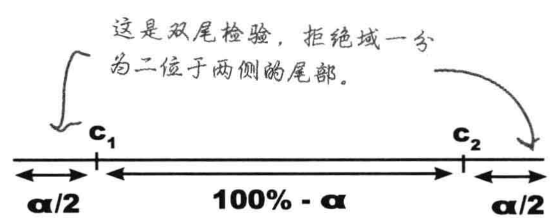
如果备择假设包含一个>符号，则使用**右尾**，此时拒绝域位于数据的高端。



我们对鼾克使用的是单尾检验，由于备择假设为$p < 0.9$，因此拒绝域位于左尾。

双尾检验

双尾检验即拒绝域一分为二位于数据集的两侧，你选择检验水平α，然后将拒绝域一分为二，并确保整个拒绝域以相应概率反映这个检验水平。两侧各占α/2，因此总和为α。



判断是否需要使用双尾检验的方法是：查看备择假设H_1，如果H_1包含一个不等号≠，则需要使用双尾检验，这是因为你要找出参数的变化，而不是增减。

对于鼾克，如果备择假设为$p \neq 0.9$，则我们应使用双尾检验，我们应该查看治愈的人数是否显著多于或显著少于90%。

第4步：求出p值

讲过拒绝域之后，我们就能进入第4步：求出P值。

p值即某个小于或者等于拒绝域方向上的一个样本数值的概率。具体求法是利用样本进行计算，然后判定样本结果是否落在假设检验的拒绝域以内。也就是说，我们通过P值确定是否该拒绝原假设。

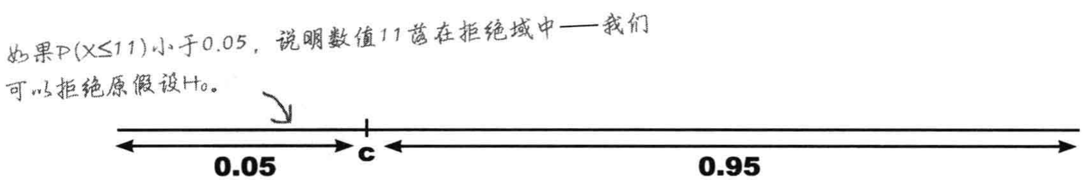
你在进行这一步 →

- 确定要进行检验的假设
- 选择检验统计量
- 确定用于做决策的拒绝域
- **求出检验统计量的p值**
- 查看样本结果是否位于拒绝域内
- 作出决策

如何求p值？

具体用哪种方法求p值取决于拒绝域和检验统计量。对于鼾克检验来说，治愈人数为11人，而拒绝域位于分布的低端，于是P值为P(X≤11)，其中X为样本中的治愈人数的分布。

由于检验的显著性水平为5%，说明如果P(X≤11)小于0.05，则数值11落在拒绝域中，这时我们可以拒绝原假设。



动动笔

我们在第2步中了解到X ~ B(15, 0.9)。那么P(X ≤ 11)等于多少？

动动笔解答

我们在第2步中了解到$X \sim B(15, 0.9)$。那么$P(X \leqslant 11)$等于多少？

$$
\begin{aligned}
P(X \le 11) &= 1 - P(X \ge 12) \\
&= 1 - ({}^{15}C_{12} \times 0.1^3 \times 0.9^{12} + {}^{15}C_{13} \times 0.1^2 \times 0.9^{13} + {}^{15}C_{14} \times 0.1 \times 0.9^{14} + 0.9^{15}) \\
&= 1 - (0.1285 + 0.2669 + 0.3432 + 0.2059) \\
&= 1 - 0.9445 \\
&= 0.0555
\end{aligned}
$$

${}^{15}C_{15}=1$也等于0.1^0，因此只需计算0.9^{15}.

我们已经求得p值

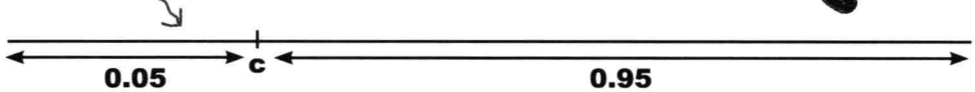
为了求得假设检验的P值，我们必须求出$P(X \le 11)$，即P值等于0.0555。



p值即为取得样本中的各种结果或取得拒绝域方向上的某些更为极端的结果的概率。

在斛克假设检验中，拒绝域位于概率分布的左尾。为了了解“治愈11位患者”这个结果是否位于拒绝域内，我们计算了$P(X \le 11)$，因为这正是取得位于左尾方向上并至少以样本结果为极值的数值的概率。

我们想了解是否“治愈11位患者”这个结果位于这个拒绝域中，因此用$P(X \le 11)$进行估计。



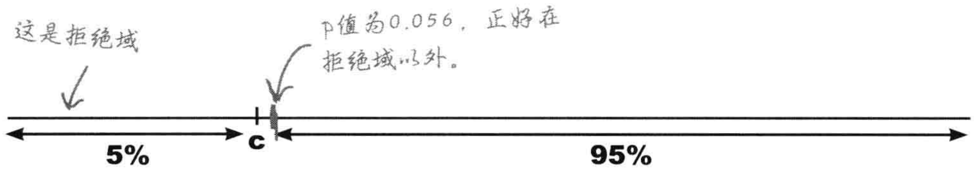
相反，假如我们的拒绝域位于概率分布的右尾，我们就需要求$P(X \ge 11)$。我们应该将更为极端的一些数值视为大于11的极值，因为这些数值本来就距离拒绝域更近。

第5步：样本结果位于拒绝域中吗？

- 确定要进行检验的假设
- 选择检验统计量
- 确定用于做决策的拒绝域
- 求出检验统计量的p值
- **查看样本结果是否位于拒绝域内** ← 你在进行这一步
- 作出决策

我们已经求出了P值，可以用它检查我们的样本结果是否落在拒绝域内。如果的确如此，则我们就有足够的证据否定制药公司的断言。

我们的拒绝域位于概率分布的左尾，所用显著性水平为5%。这意味着，如果P值小于0.05，就能拒绝原假设。由于我们的P值为0.0555，因此样本中用鼾克治愈的患者数不在拒绝域内。



第6步：作出决策

- 确定要进行检验的假设
- 选择检验统计量
- 确定用于做决策的拒绝域
- 求出检验统计量的p值
- 查看样本结果是否位于拒绝域内
- **作出决策** ← 你在进行这一步

我们已经进入假设检验的最后一步：决定接受原假设，还是拒绝原假设而改用备择假设。

因为假设检验的P值落在检验的拒绝域以外，因此，没有充分的证据可以拒绝原假设。所以：

我们接受制药公司的断言



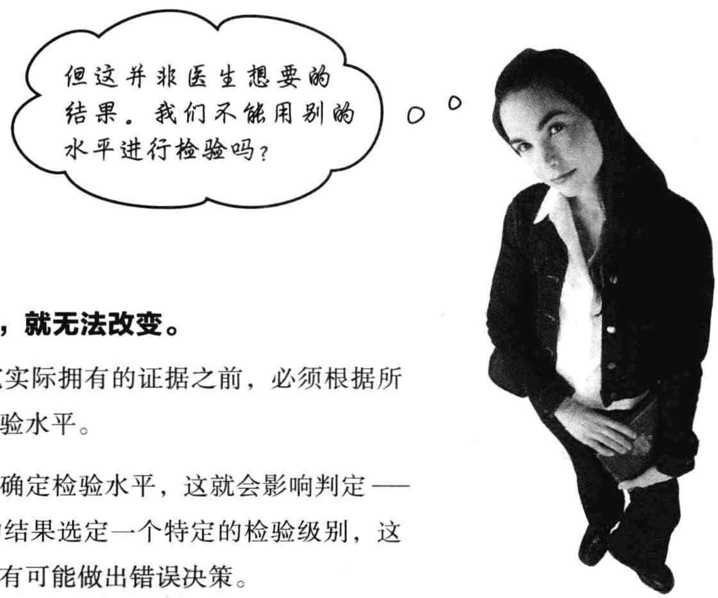
我们前面做了哪些工作？

让我们总结一下前面的工作。

首先，我们取用制药公司的断言——医生对此断言有疑虑。我们将这些断言作为假设检验的基础，形成一个原假设：患者的治愈概率为0.9，随后将这个概率应用于医生样本的人数。

然后，我们决定以5%的检验水平进行检验，检验中使用了医生的样本治愈率。我们计算了有11位或11位以下患者得到治愈的概率，然后检查这个概率是否低于5%，也就是0.05。换句话说，我们计算了等于这个极值或比这个极值更极端的数值的概率。

最后，我们求出：当检验水平为5%时，没有足够的证据可以否定制药公司的断言。



一旦确定了检验的显著性水平，就无法改变。

检验必须绝对公正。因此在研究实际拥有的证据之前，必须根据所需要的证据水平决定所需要的检验水平。

如果打算先看证据是否充分，再确定检验水平，这就会影响判定——你可能会忍不住按照心中想要的结果选定一个特定的检验级别，这就会令检验结果发生偏倚，于是有可能做出错误决策。

要 点

- 进行假设检验即选定一个断言，然后借助统计证据对其进行检验。
- 所检验的断言被称为原假设，用H_0表示。除非有有力的证据证明断言不正确，否则就接受断言。
- 备择假设即在有充分证据拒绝原假设H_0的情况下将接受的假设，用H_1表示。
- 检验统计量即用于对假设进行检验的统计量，是与检验具有最密切关系的统计量。选择检验统计量的时候，你假定H_0为真。
- 显著性水平用α表示，它表示你希望在观察结果的不可能程度达到多大时拒绝H_0。
- 拒绝域为一组数值，代表可用于否定原假设的最极端证据。选择拒绝域时，需考虑显著性水平，还要考虑用单尾还是双尾进行检验。
- 单尾检验的拒绝域位于数据的左侧或右侧，双尾检验的数据一分为二位于数距的两侧。可根据备择假设选择尾部。
- P值即取得样本结果或取得拒绝域方向上的更极端结果的概率。
- 如果P值位于拒绝域中，则有充足的理由拒绝原假设；如果P值位于拒绝域以外，则没有充足的证据。

世上没有傻问题

问： 一般可用哪种显著性水平进行检验？

答： 这完全取决于你希望以多大力度的证据拒绝原假设。你越想增大证据力度，显著性水平必须越小。

最常用的显著性水平为5%，不过有时也会用到1%的显著性水平。用1%的水平进行检验意味着证据力度大于5%的水平。

问： 显著性水平与置信区间的置信水平有共同之处吗？

答： 有，有不少共同之处。在为总体参数构建置信区间时，你希望对“总体参数位于两个限值之间”这一结果具有一定的置信度，例如，如果置信水平为95%，则说明总体参数位于两个限值之间的概率为0.95。

显著性水平反映了数值将位于某个限值以外的概率。例如显著性水平为5%意味着拒绝域的概率必须为0.05。



如果样本增大会怎么样？

前面讲过，医生仅以15人为样本进行了试验，以这个样本为依据得出的证据不足以否定制药公司的断言。

有可能样本不够大，这才无法得出正确的结果。如果医生使用一个大一点儿的样本，可能会得出更可靠的结果。

下面是医生的新试验结果：

是否治愈？	是	否
频数	80	20



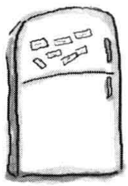


我们希望确定：新数据是否会使检验结果发生变化。

让我们再进行一次假设检验，这一次用一个更大的样本。

动动脑

新问题的原假设是什么？备择假设是什么？



假设检验磁贴

现在该进行另一个假设检验了，这需要执行一系列步骤。你还记得这些步骤的顺序吗？请将磁贴按正确顺序放好。

作出决策

选择检验统计量

确定要进行检验的假设

确定用于做决策的拒绝域

求出检验统计量的p值

查看检验统计量是否位于拒绝域内



假设磁贴解答

现在该进行另一个假设检验了，这需要执行一系列步骤。你还记得这些步骤的顺序吗？请将磁贴按正确顺序放好。

确定要进行检验的假设

选择检验统计量

确定用于做决策的拒绝域

求出检验统计量的p值

查看检验统计量是否位于拒绝域内

作出决策

让我们再进行一次假设检验

确定要进行检验的假设
选择检验统计量
确定用于做决策的拒绝域
求出检验统计量的p值
查看样本结果是否位于拒绝域内
作出决策

医生对于制药公司的断言仍有疑虑。

让我们根据新数据进行一次假设检验。

第1步：确定假设

我们首先需要确定鼾克的原假设和备择假设。提醒一下：原假设即我们正在进行检验的断言，备择假设则是在没有充分证据拒绝原假设的情况下接受的假设。

那么原假设是什么？备择假设又是什么？

还是老问题

在上一次检验中，我们采用制药公司的断言，以此为基础形成原假设。我们现在要对同样的断言进行检验，因此原假设还是老样子，已知：

$$H_0 : p = 0.9$$

备择假设也是老样子。如果有有力的证据否定制药公司的断言，则我们将接受“药物的患者治愈率低于90%”这一说法，于是备择假设为：

$$H_1 : p < 0.9$$



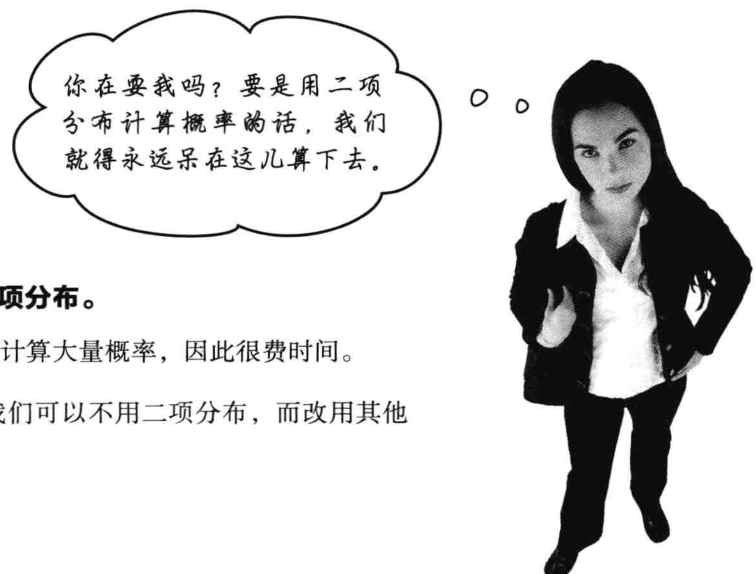
第2步：选择检验统计量

- 确定要进行检验的假设
- **选择检验统计量**
- 确定用于做决策的拒绝域
- 求出检验统计量的p值
- 查看样本结果是否位于拒绝域内
- 作出决策

如上所述，第2步是选择检验统计量，即需要找出某个统计量，以便对假设进行检验。

在前一个假设检验中，我们通过观察样本的成功数目以及结果的显著性进行检验。我们用二项分布求出了一个至少以样本数值为极值的概率。换句话说，我们用检验统计量$X \sim B(15, 0.9)$检验$P(X \leq 11)$是否小于显著性水平0.05。

这一次，样本中的患者数是100，要检验的断言还是老样子——治愈某位患者的概率为0.9，即我们的新检验统计量为$X \sim B(100, 0.9)$。



我们可以用另一种分布代替二项分布。

用二项分布解决这一类问题需要计算大量概率，因此很费时间。

幸运的是，还有另一种方法。我们可以不用二项分布，而改用其他分布。

动动脑

你能用哪种概率分布近似$X \sim B(100, 0.9)$?



练习

为了能够最大限度地发挥假设检验的作用，你需要了解各种变量和参数的分布情况。在下列情况下，你会用哪种分布求概率？

提示：本书前文已经对这些情况进行过讲解。若有疑难，请参看前文。

1. $X \sim B(n, p)$。如果n很大，$np > 5$且$nq > 5$，你会用哪种概率分布进行近似？

2. $X \sim N(\mu, \sigma^2)$。已知μ和σ^2的数值，$\overline{X}$符合哪种分布？

3. $X \sim N(\mu, \sigma^2)$。已知μ，但不知道σ^2的大小，样本很大。假如数据已知，那么$\overline{X}$符合什么分布？

4. $X \sim N(\mu, \sigma^2)$。已知μ，但不知道σ^2的大小，样本很小。假如数据已知，那么$\overline{X}$符合什么分布？



为了能够最大限度地发挥假设检验的作用，你需要了解各种变量和参数的分布情况。在下列情况下，你会用哪种分布求概率？

1. $X \sim B(n, p)$。如果n很大，$np > 5$且$nq > 5$，你会用哪种概率分布进行近似？

如果n很大，则我们可以用正态分布近似$X \sim B(n, p)$。由于$E(X) = np$，$Var(X) = npq$，于是可以用$X \sim N(np, npq)$，其中假定$np > 5$，$nq > 5$。

2. $X \sim N(\mu, \sigma^2)$。已知μ和σ^2的数值，$\overline{X}$符合哪种分布？

如果我们知道σ^2的数值，则$\overline{X} \sim N(\mu, \sigma^2/n)$。

3. $X \sim N(\mu, \sigma^2)$。已知μ，但不知道σ^2的大小，样本很大。假如数据已知，那么$\overline{X}$符合什么分布？

如果我们不知道σ^2的数值，则用s^2进行估计，$\overline{X} \sim N(\mu, s^2/n)$.

4. $X \sim N(\mu, \sigma^2)$。已知μ，但不知道σ^2的大小，样本很小。假如数据已知，那么$\overline{X}$符合什么分布？

如果我们不知道σ^2的数值，则用s^2进行估计，如果样本很小，则使用t分布$T \sim t(n-1)$，其中

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

在我们的检验统计中用正态分布近似二项分布

我们照样需要找到一个能用于进行假设检验的检验统计量。由于样本数量很大，使用二项分布将会费时费力。

样本中有100名患者，而按照制药公司的说法，成功比例为0.9。这就是说，成功数目服从二项分布，其中n = 100，P = 0.9。

由于n很大，且np和nq都大于5，我们就用X~N(np, npq)作为检验统计量，其中X为成功治愈的患者的数目。即我们能够用

$$X \sim N(90, 9)$$

由于n很大，且np > 5以及nq很大，因此我们可以用这个分布。

近似我们所需要的任何概率。

经过标准化，得到：

$$Z = \frac{X - 90}{\sqrt{9}}$$

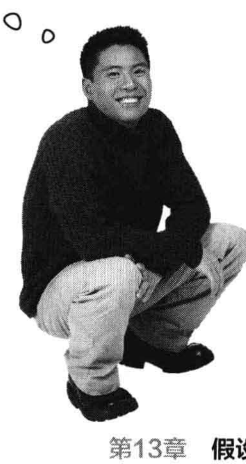
对X ~ N(90, 9)进行标准化。

$$= \frac{X - 90}{3}$$

也就是说，我们的检验统计量可以是：

$$Z = \frac{X - 90}{3} \qquad Z \sim N(0, 1)$$

X是治愈患者的数目，在我们的实例中，这个数目是80。



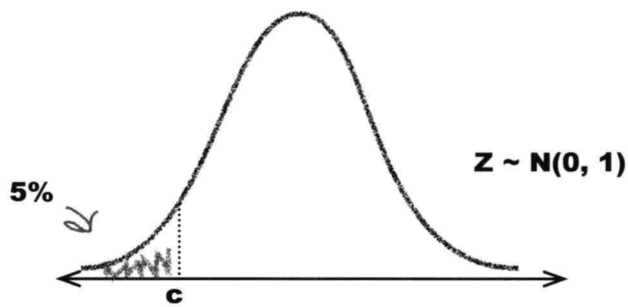
你用检验统计量计算概率——该概率可以当作证据。

这就是说，我们将Z作为检验统计量——因为通过它可以轻松查出概率，进而了解在以制药公司断言为前提的情况下，我们的样本结果的不可能程度如何。我们将80代入X，这样就能求出治愈人数为80或80以下的概率。

第3步：求出拒绝域

- 确定要进行检验的假设
- 选择检验统计量
- **确定用于做决策的拒绝域**
- 求出检验统计量的p值
- 查看样本结果是否位于拒绝域内
- 作出决策

有了检验统计量之后，还需要求拒绝域。由于我们的备择假设为$p < 0.9$，这表明拒绝域位于左尾，这和前面是一样的。拒绝域还取决于检验的显著性水平，让我们选择和前面一样的显著性水平，即以5%水平进行检验。

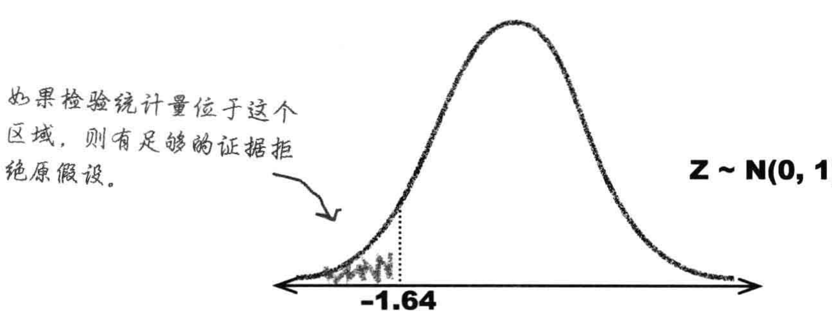


由于我们的检验统计量符合标准正态分布，于是可以用概率表查出临界值c。临界值即具有足够证据拒绝原假设和不具有足够证据拒绝原假设这两种情况的分界值。

由于我们的显著性水平为5%，于是临界值c等于令$P(Z < c)=0.05$的数值。在概率表中查找0.05，得到c的数值为−1.64，即：

$$P(Z < -1.64) = 0.05$$

这说明只要检验统计量小于−1.64，我们就有足够的证据拒绝原假设。





你觉得自己能完成其余假设检验步骤吗？看看能否求出下列结果：

第4步：求p值

拒绝域位于分布的左尾，治愈人数为80人，Z = (X - 90)/3，利用这些条件求出p值。

第5步：查看检验统计量是否位于拒绝域内

别忘了：假设检验的显著性水平为5%。

第6步：作出决策

根据证据，你接受还是拒绝原假设？



你觉得自己能完成其余假设检验步骤吗？看看能否求出下列结果：

第4步：求p值

拒绝域位于分布的左尾，治愈人数为80人，Z = (X – 90)/3，利用这些条件求出P值。

让我们先求80的标准分。

$$z = (80 - 90)/3$$
$$= -10/3$$
$$= -3.33$$

p值算法为$P(Z < z) = P(Z < -3.33)$，查找概率表，得：

p值 = 0.0004

第5步：查看检验统计量是否位于拒绝域内

别忘了：假设检验的显著性水平为5%。

如果P值小于0.05，则检验统计量位于拒绝域中。由于P值等于0.0004，说明检验统计量位于拒绝域中。

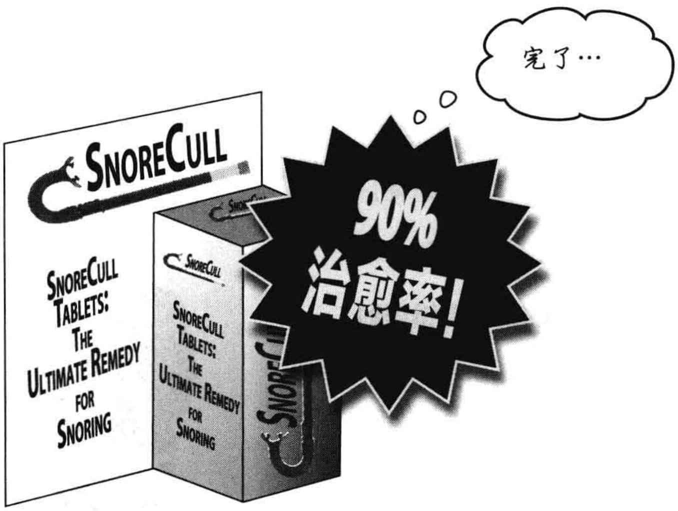
第6步：作出决策

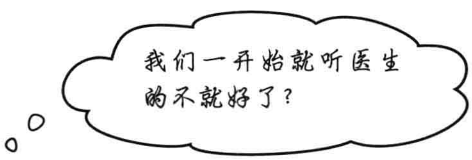
根据证据，你接受还是拒绝原假设？

由于检验统计量位于假设检验的拒绝域中，说明在5%显著性水平的情况下，我们有足够的证据拒绝原假设。

鼾克未通过检验

在对鼾克进行的这一次检验中，有足够的证据证明可以拒绝原假设，这说明我们可以否定制药公司的断言。







假设检验需要证据。

进行假设检验时，你选取一个断言，然后对其进行试验。只有在有足够证据反驳这个断言时，你才能否定这个断言。这意味着检验是公正的，因为你做决策的唯一依据就是是否有充分证据。

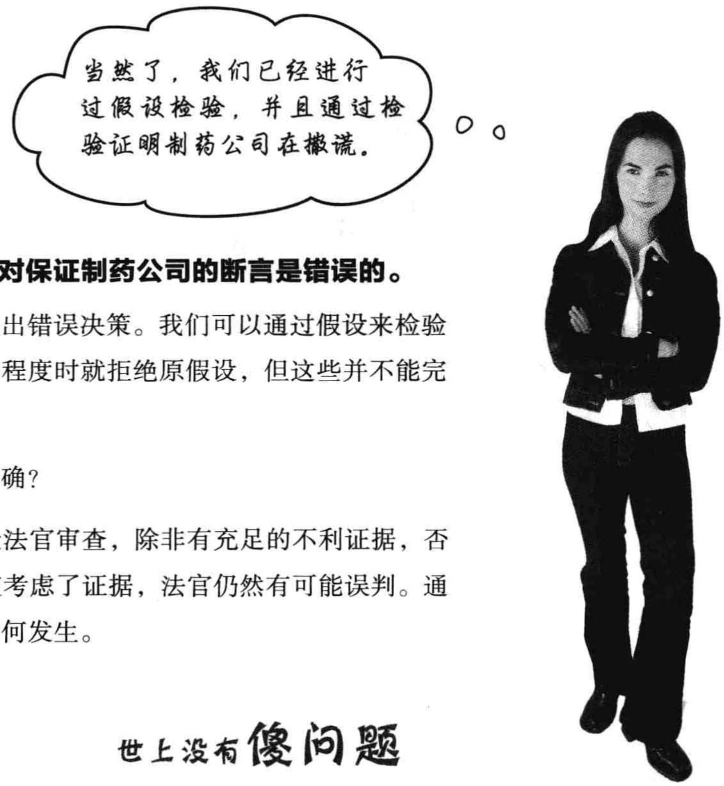
如果我们一开始就接受医生的观点，就不会妥当地考虑证据。我们会在不考虑结果是否只能解释为偶然的情况下作出决策，而现在呢，我们有足够的证据表明，样本结果足以合理地拒绝原假设。这些结果具有**统计显著性**，因为它们不可能是偶然发生的。

这能保证制药公司的断言是错误的吗？

可能出现错误

前面讲到在假设检验中如何将样本结果作为证据，如果证据足够有力，则我们用这些证据合理地否定原假设。

我们已经发现有足够证据证明制药公司的断言是错误的，但是，能对此做出保证吗？



即使证据很有力，我们也无法绝对保证制药公司的断言是错误的。

说是说不可能，但我们仍然可能做出错误决策。我们可以通过假设来检验证据，可以规定在确定性达到何种程度时就拒绝原假设，但这些并不能完全保证我们的决策是正确的。

问题是，我们如何确定决策是否正确？

进行假设检验有点儿像让囚犯接受法官审查，除非有充足的不利证据，否则法官假定囚犯无罪，但是，即使考虑了证据，法官仍然有可能误判。通过下一页的练习，你将明白误判如何发生。

世上没有傻问题

问： 在进行假设检验的时候，我们怎么会做出错误决策呢？我们做假设检验不就是为了确保不判错吗？

答： 在进行假设检验的时候，你只能根据手头拥有的证据作决策，证据来源于样本，因此，如果样本有偏，那么你就会根据有偏数据做出错误决策。

问： 我曾经听人说起过“显著性检验”，这是什么？

答： 有些人把假设检验称为显著性检验，这是因为你是按照某种显著性水平进行检验的。

动动笔

一个囚犯正在因犯罪行为接受审判，你是法官。法官的任务是假定囚犯无罪，但是，假如有足够证据证明囚犯有罪，则需宣判囚犯有罪。

1. 这个试验的原假设是什么？

2. 备择假设是什么？

3. 在什么情况下，法官做出正确判决？

4. 在什么情况下，法官会做出错误判决？

动动笔解答

一个囚犯正在因犯罪行为接受审判，你是法官。法官的任务是假定囚犯无罪，但是，假如有足够证据证明囚犯有罪，则需宣判囚犯有罪。

1. 这个试验的原假设是什么？

原假设是：囚犯无罪。除非有反面证据，否则我们必须如此假定。

2. 备择假设是什么？

备择假设是：囚犯有罪。也就是说，如果有充分证据证明囚犯并非无罪，则我们接受囚犯有罪这一说法，并进行宣判。

3. 在什么情况下，法官做出正确判决？

如下行事可进行正确判决：

囚犯无罪，且我们发现他无罪。

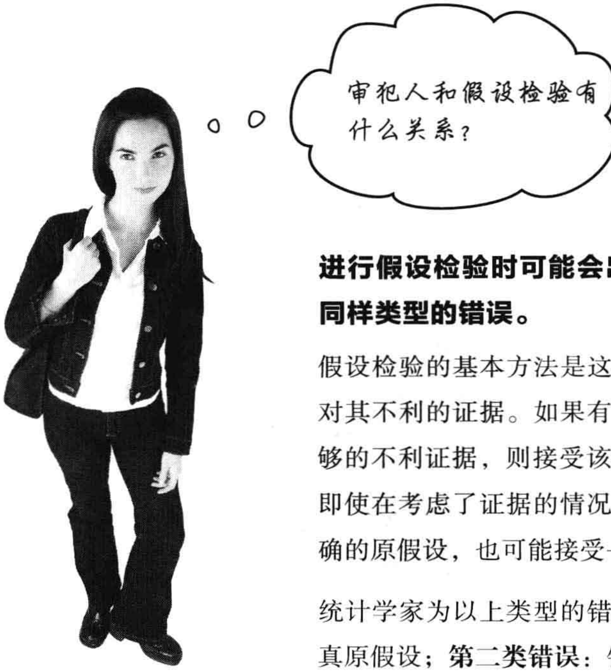
囚犯有罪，且我们发现他有罪。

4. 在什么情况下，法官会做出错误判决？

如下行事可做出错误判决：

囚犯无罪，而我们发现他有罪。

囚犯有罪，而我们发现他无罪。



进行假设检验时可能会出现的错误与审判罪犯时可能会犯的错误是同样类型的错误。

假设检验的基本方法是这样的：选取一个断言，对其进行检验——评估对其不利的证据。如果有足够的不利证据，则否定该断言；如果没有足够的不利证据，则接受该断言。你可能会正确地接受或拒绝原假设，但即使在考虑了证据的情况下，仍然有可能犯错误。你可能会拒绝一个正确的原假设，也可能接受一个实质上错误的原假设。

统计学家为以上类型的错误给出了专用名称。**第一类错误**：错误地拒绝真原假设；**第二类错误**：错误地接受假原假设。

假设检验的**功效**即你正确地拒绝一个假原假设的概率。

假设检验决策

实际情况

	接受H_0	**拒绝H_0**
H_0真	✓	第一类错误
H_0假	第二类错误	✓

这些都是错误

这给出检验的功效。

动动脑

你认为我们该如何求出发生第一类错误的概率？如何求出发生第二类错误的概率？

让我们从第一类错误讲起

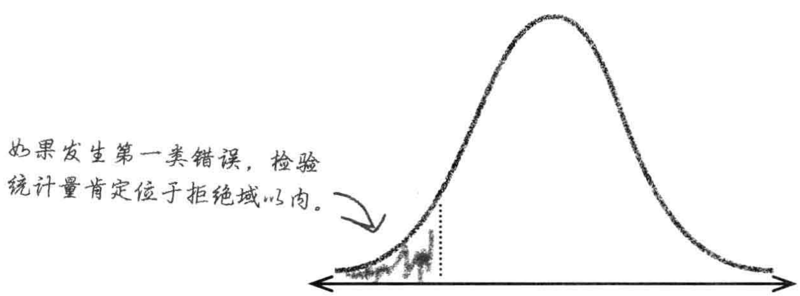
第一类错误即在原假设实际为正确的情况下拒绝原假设的后果。就像审判囚犯，发现其有罪，但实际上他却无罪。

若H_0实际上是正确的，你却拒绝了，这就发生了第一类错误。



发生第一类错误的概率是多大？

如果发生第一类错误，那么一定是拒绝了原假设。拒绝原假设的前提是：样本结果必须位于拒绝域以内。



发生第一类错误的概率等于你的结果位于拒绝域以内的概率。由于拒绝域由检验水平决定，说明如果检验的显著性水平为α，则发生第一类错误的概率必须也等于α。

即：

$$P(\text{第一类错误}) = \alpha$$

其中α为检验的显著性水平。

再谈第二类错误

当原假设实际为错误假设时，如果你接受原假设，则发生第二类错误。
这就像对一个囚犯进行审判，发现其无罪，但实际上他是有罪的。



当H_0为错误假设而你接受该假设时，则发生第二类错误。

发生第二类错误的概率通常用希腊字母β表示。

P(第二类错误) = β

如何求β？

求第二类错误的概率要比求第一类错误的概率难得多。下面是相关步骤，我们将在下一页讲解执行过程。

1. **检查是否拥有H_1的特定数值。**
   没有这个数值则无法计算第二类错误概率。
2. **求检验拒绝域以外的数值范围。**
   如果检验统计量已经标准化，则该数值范围要进行逆标准化。
3. **假定H_1为真，求得到这些数值的概率。**
   也就是说，我们要求出得到拒绝域以外的数值的概率，但这一次用H_1而不是H_0对检验统计量进行描述。.

发现斛克检验的错误

让我们看看是否能求出斛克假设检验发生第一类错误和第二类错误的概率。

$$Z = \frac{X - 90}{3}$$

其中X为样本中的治愈患者数。检验的显著性水平为5%。

让我们从第一类错误算起

第一类错误即在原假设实际上为真时却拒绝原假设所引起的错误，发生这种错误的概率与假设检验的显著性水平相等，即：

P(第一类错误) = 0.05

这就是在“治愈率为90%”这个原假设为真时却拒绝原假设的概率。

第二类错误如何计算？

第一类错误即在备择假设为真时却接受原假设所引起的错误，只有在H_1规定了唯一特定值时我们才能计算这个错误，因此让我们使用备择假设P = 0.8，因为这个值是医生样本的成功比例。于是我们的假设为：

H_0: P = 0.9

H_1: P = 0.8

这一次，我们用的是H_1: P = 0.8而不是H_1: P < 0.8，只有在备择假设具有唯一特定值时才能计算第二类错误的发生概率。

H_1必须规定一个确切的P值，因为只有这样我们才能利用它计算概率。如果我们使用备择假设P < 0.9，那么无法利用它计算发生第二类错误的概率。

为了能用备择假设概率分布查找概率，我们需要一个确切的P值。



放轻松

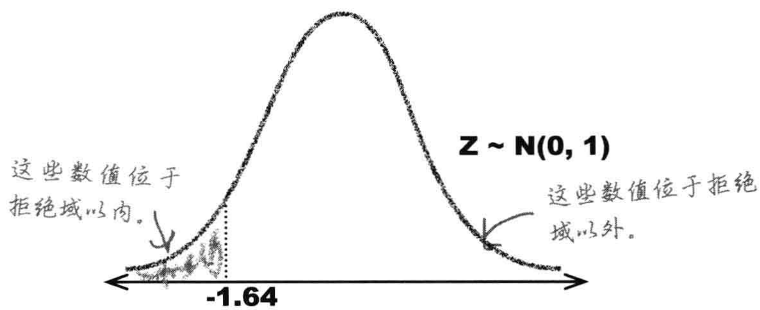
如果在考试中需要计算发生第二类错误的概率，题目会告诉你H_1是多少。

这就是说不需要你自己确定备择假设。如果需要计算这一类错误，这将是已知条件。.

我们需要求数值范围

既然备择假设H_1有了一个特定的P值，我们就能进入下一步了。我们需要求出位于假设检验拒绝域以外的X值。

回头查阅548页，我们会看到检验的拒绝域由Z< -1.64给出，即，P(Z < -1.64) = 0.05。这说明拒绝域以外的数值由Z ≥ -1.64给出。



经过逆标准化，得到：

$$\frac{X - 90}{3} \geq -1.64$$

$$X - 90 \geq -1.64 \times 3$$

$$X \geq -4.92 + 90$$

$$X \geq 85.08$$

即，如果斛克的治愈人数为85.08或更多，则我们就会接受原假设。

最后，我们需要假定H_1为真，算出P(X ≥ 85.08)，这样我们就能算出在H_1实际上为真的情况下接受原假设的概率。由于我们使用正态分布近似X，于是需要使用的概率分布为X ~ N(np, npq)，其中n = 100，p = 0.8，得到X ~ N(80, 16)。

$$X \sim N(80, 16)$$

这说明，如果我们算出P(X ≥ 85.08)，其中X ~ N(80, 16)，我们就能求出发生第二类错误的概率。

该概率的计算方法与其他正态分布概率的算法相同：求出标准分，然后在标准正态分布表中查找数值。

求P(第二类错误)

通过计算P(X ≥ 85.08)，其中X ~ N(80, 16)，我们可以求出发生第二类错误的概率。让我们先求85.08的标准分。

这是常用的计算标准分的方法：减去期望，然后除以标准差。

$$z = \frac{85.08 - 80}{\sqrt{16}}$$

$$= \frac{5.08}{4}$$

$$= 1.27$$

即，为了求P(X ≥ 85.08)，我们需要使用标准概率表求出P(Z ≥ 1.27)。

$$P(Z \geq 1.27) = 1 - P(Z < 1.27)$$

$$= 1 - 0.8980$$

$$= 0.102$$

即：

P(第二类错误) = 0.102

这就是在实际上能治愈80%患者的情况下，接受"能治愈90%患者"这个原假设的概率。

世上没有傻问题

问： 求P(第二类错误)为什么比求 P(第一类错误) 难这么多?

答： 这是由其定义决定的。第一类错误是错误拒绝原假设所引起的结果；发生这类错误的概率等于α——检验的显著性水平。

第二类错误是在备择假设实际上为真的情况下接受原假设所引起的结果，为了求出发生这一类错误的概率，你首先需要求出样本中的表明你接受原假设的数值范围。在求出这些数值之后，还需要计算在假设H_1为真的情况下取得这些数值的概率。

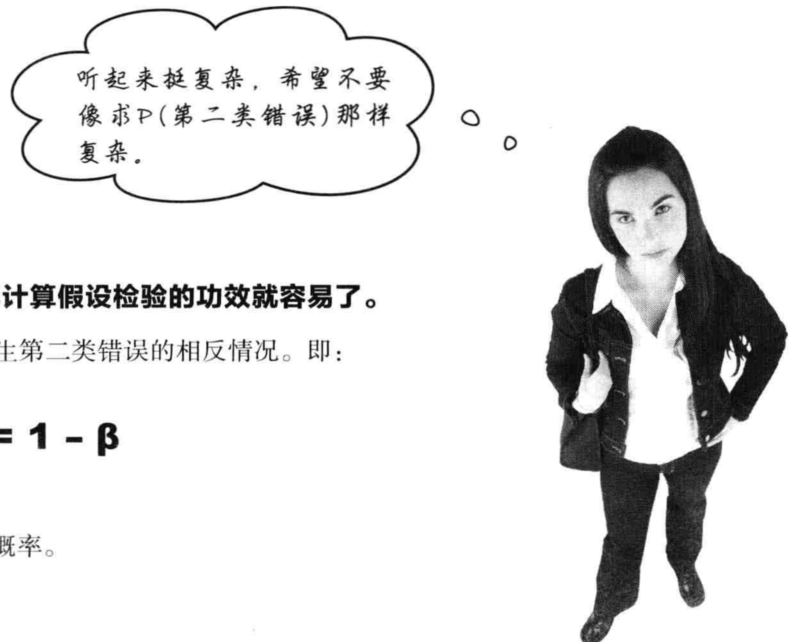
问： 每当我想求发生第二类错误的概率时，都要用正态分布吗?

答： 所用概率分布取决于检验统计量。在我们的例子中，检验统计量符合正态分布，因此用正态分布求P(第二类错误)。如果检验统计量符合其他分布，例如泊松分布，则应该用泊松分布。

认识功效

前面讲到进行假设检验时所发生的各种错误的概率，还有一事尚未谈及：功效。

假设检验的**功效**也是一种概率——在H_0为假的情况下拒绝H_0的概率。也就是说，这是我们做出正确决策而拒绝H_0的概率。



只要求出P(第二类错误)，再计算假设检验的功效就容易了。

在H_0为假时拒绝H_0其实就是发生第二类错误的相反情况。即：

$$\text{功效} = 1 - \beta$$

其中β等于发生第二类错误的概率。

鼾克假设检验的功效是多少？

我们已经求得第二类错误的概率为0.102，通过下式可算得鼾克假设检验的功效：

$$\begin{aligned}\text{功效} &= 1 - P(\text{第二类错误}) \\ &= 1 - 0.102 \\ &= 0.898\end{aligned}$$

即，鼾克假设检验的功效为0.898，因此我们做出正确决策而拒绝原假设的概率为0.898。

医生开心了

你在本章进行了两次假设检验，证实有充分证据否定制药厂的断言。你能够阐明，根据医生的样本，有足够的证据证明鼾克无法治愈90%的鼻鼾患者，而制药厂却断言可以做到。





不过事情还没有到此结束

请接着看下去，我们将介绍其他可供使用的假设检验。肥蛋赌场见……

练习
(上)

制药公司和他们的止咳糖浆制造厂发生了争议，厂方说注入药瓶的糖浆量符合正态分布 $X \sim N(355, 25)$，其中X是量得的每瓶糖浆容量，单位mL。制药公司用大样本进行了检验，发现100瓶糖浆的平均容量为356.5mL。请以1%的显著性水平检验厂方给出的均值假设，与此相对的另一说法是每瓶糖浆的容量均值大于355mL。

第1步：确定要进行检验的假设。原假设是什么？备择假设是什么？

第2步：选择检验统计量。

提示：你的假设涉及到均值，那么$\overline{X}$符合什么分布？如何进行标准化？

第3步：决定用于做决策的拒绝域。拒绝域位于分布的左尾还是右尾？显著性水平是多少？



练习解答（上）

制药公司和他们的止咳糖浆制造厂发生了争议，厂方说注入药瓶的糖浆量符合正态分布X ~ N(355, 25)，其中X是量得的每瓶糖浆容量，单位mL。制药公司用大样本进行了检验，发现100瓶糖浆的平均容量为356.5mL。请以1%的显著性水平检验厂方给出的均值假设，与此相对的另一说法是每瓶糖浆的容量均值大于355mL。

第1步：确定要进行检验的假设。原假设是什么？备择假设是什么？

我们想检验每瓶糖浆的容量均值是否如厂方所述为355mL，因此：

H_0: $\mu = 355$

H_1: $\mu > 355$

第2步：选择检验统计量。

$\overline{X} \sim N(\mu, \sigma^2/n)$，因此根据原假设得知：$\overline{X} \sim N(355, 25/100)$ 或 $\overline{X} \sim N(355, 0.25)$。

对此进行标准化，得到：

$$Z = \frac{\overline{X} - 355}{\sqrt{0.25}}$$

$$= \frac{\overline{X} - 355}{0.5}$$

第3步：决定用于做决策的拒绝域。拒绝域位于分布的左尾还是右尾？显著性水平是多少？

备择假设为 $\mu > 355$，即拒绝域位于右尾。我们想以1%的显著性水平进行检验，因此拒绝域由$P(Z > c) = 0.01$决定。利用概率表，得到：$c = 2.32$。即拒绝域由$Z > 2.32$确定。



练习
（下）

继续前面的练习：这是假设检验的后三步。你能得出什么结论？

第4步：求假设检验的p值。使用分布Z = ($\overline{X}$ - 355)/0.5，即样本糖浆的容量均值，记住，这一次你需要查看检验统计量是否位于分布的右尾，因为这正是拒绝域所在位置。

第5步：查看样本结果是否位于拒绝域以内。记住：检验的显著性水平是1%。

第6步：作出决策。是否有足够的证据拒绝显著性水平为1%的原假设？



练习解答（下）

继续前面的练习：这是假设检验的后三步。你能得出什么结论？

第4步：求假设检验的p值。使用分布$Z = (\overline{X} - 355)/0.5$，即样本糖浆的容量均值，记住，这一次你需要查看检验统计量是否位于分布的右尾，因为这正是拒绝域所在位置。

$$Z = (\overline{X} - 355)/0.5$$

$$= (356.5 - 355)/0.5$$

$$= 1.5/0.5$$

$$= 3$$

由于拒绝域位于右尾，因此检验的p值由$P(Z > 3)$决定，查概率表，得到：

P值 $= 0.0013$

第5步：查看样本结果是否位于拒绝域以内。记住：检验的显著性水平是1%。

P值0.0013小于显著性水平0.01，这表明样本结果位于拒绝域以内。

第6步：作出决策。是否有足够的证据拒绝显著性水平为1%的原假设？

由于样本结果位于拒绝域以内，有充分的证据拒绝原假设。我们可以接受备择假设：$\mu > 355$ ml。

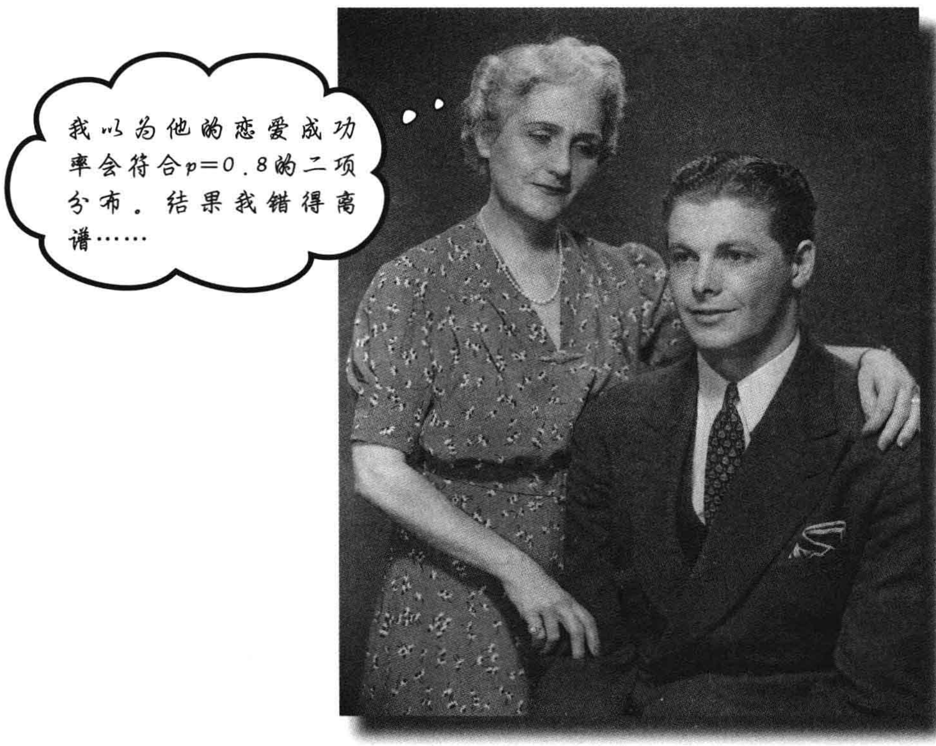


要点

- 第一类错误即在原假设正确时却拒绝原假设。发生第一类错误的概率为α —— 即检验的显著性水平。
- 第二类错误即在原假设错误时却接受原假设。发生第二类错误的概率用β表示。
- 为了求出β，备择假设必须为一个特定数值。于是你求出检验拒绝域以外的数值范围，然后求出以H_1为条件得到这个数值范围的概率。

14 χ^2分布

继续探讨……



有时候事实与期望并不相符。

当以一种特定的概率分布为某种情况建模时，对于事物的长期可能结果，你有十分清晰的想法。可如果**期望与事实**存在差别呢？你该如何判断？—— 这些偏差是正常波动，还是说明概率模型存在问题？本章将讲解如何利用χ^2分布**分析结果**，排除**可疑结果**。

肥蛋赌场可能有麻烦

肥蛋赌场惯于从赌客身上捞钱，不过，这个星期它碰到了问题—— 老虎机总是出头奖，轮盘总是停在12位，骰子老是不称手，有一张赌二十一点的牌桌上出现了太多赢家。

赌场再这么赔下去就撑不住了，肥蛋老板怀疑有人动了手脚，他需要**你**帮他探明究竟。



让我们从老虎机开始

前面已经讲过，肥蛋赌场有一大排亮闪闪的老虎机，只等着大家去赌。问题是，人们不仅赌个不停——而且赢个不停。

下面是某台老虎机的期望概率分布，其中X代表每一局游戏的净收益：

每局2美元，如果什么也赢不到的话，你就损失2美元。

如果中了头奖，净收益就是98美元。

x	-2	23	48	73	98
P(X = x)	0.977	0.008	0.008	0.006	0.001

赌场搜集了一些统计数据，给出了人们获得某种收益的次数。下面是观察到的每局净收益的频数：

频数指出每种收益的发生次数。

x	-2	23	48	73	98
频数	965	10	9	9	7

动动笔

我们需要将每个x值的实际频数与根据概率分布得出的期望频数进行比较。请填写下表，看出什么了吗？

观察频数即实际发生的频数。

提示：总观察频数为1,000，将几个观察频数相加即可得到这个数值。请使用概率分布算出期望频数。

x	观察频数	期望频数
-2	965	977
23	10	
48	9	
73	9	
98	7	

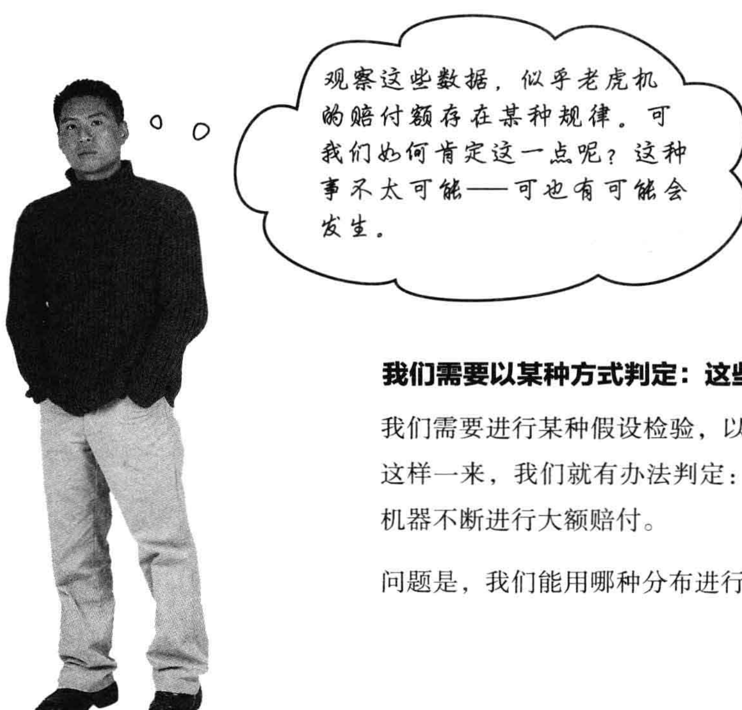
动动笔解答

我们需要将每个x值的实际频数与根据概率分布得出的期望频数进行比较。请填写下表，看出什么了吗？

x	观察频数	期望频数
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

用每种结果的概率乘以总频数1000，可得期望频数。

你根据概率分布得出的期望赢取头奖人数与实际赢取头奖人数之间有差别，但我们不知道这些差别的显著程度。



我们需要以某种方式判定：这些结果能否说明老虎机受到操纵。

我们需要进行某种假设检验，以此检验观察频数和期望频数之间的差别。这样一来，我们就有办法判定：老虎机是否被人动过手脚——以致这些机器不断进行大额赔付。

问题是，我们能用哪种分布进行这项假设检验？

用X^2检验评估差异

有一种概率分布正合我们的心意 —— χ^2分布，χ读作“卡”，是希腊字母chi的大写。这种分布通过一个检验统计量来比较期望结果和实际结果之间的差别，然后得出观察频数极值的发生概率。

让我们先求检验统计量。为此，首先画一张表，填入相应问题的观察频数和期望频数，然后，用观察频数和期望频数计算下列统计量，其中O代表观察频数，E代表期望频数。

O代表观察频数，E代表期望频数。

$$X^2 = \sum \frac{(O - E)^2}{E}$$

即，对于概率分布中的每一个概率，取期望频数和实际频数的差，求差的平方数，再除以期望频数，然后将所有结果相加。

那么老虎机问题的检验统计量是多少？

动动笔

用在上一页算出的肥蛋赌场老虎机观察频数和期望频数表计算检验统计量。看结果如何？

数值小说明什么？数值大说明什么？

动动笔解答

用在上一页算出的肥蛋赌场老虎机观察频数和期望频数表计算检验统计量。看结果如何？

数值小说明什么？数值大说明什么？

$$
\begin{aligned}
X^2 &= (965 - 977)^2/977 + (10 - 8)^2/8 + (9 - 8)^2/8 + (9 - 6)^2/6 + (7 - 1)^2/1 \\
&= (-12)^2/977 + 2^2/8 + 1^2/8 + 3^2/6 + 6^2 \\
&= 144/977 + 4/8 + 1/8 + 9/6 + 36 \\
&= 0.147 + 0.5 + 0.125 + 1.5 + 36 \\
&= 38.272
\end{aligned}
$$

如果X^2值很小，说明观察频数和期望频数之间的差别不显著；X^2越大，差别越显著。

检验统计量代表什么？

检验统计量χ^2提供了一种对观察频数和期望频数之间的差异进行量度的办法。χ^2的数值越小，观察频数和期望频数之间的总差值越小。

除数E为期望频数，于是所得结果与期望频数成反比例。

$$X^2 = \sum \frac{(O - E)^2}{E}$$

O和E之间的差值越小，X^2越小。

以E为除数，令差值与期望频数成比例。

X^2大到什么程度才算得上显著呢？—— 我们需要指出：在什么情况下才能十分肯定地判定老虎机出了问题 —— 而且这个问题已经超出了“合理偶然性”的范围。

为此我们需要讲讲**X^2分布**。

X^2分布的两个主要用途

X^2概率分布主要用于检查实际结果与期望结果之间何时存在显著差别，该概率分布使用前面讲到的检验统计量X^2进行检验。

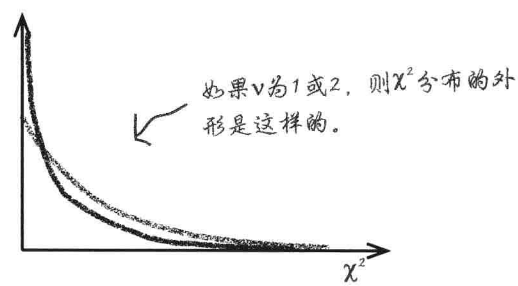
X^2分布有两个主要用途。

第一是用于检验**拟合优度**，也就是可以检验一组给定的数据与指定分布的吻合程度。例如，可以用它检验老虎机收益的观察频率与我们所期望的分布的吻合程度。

X^2分布的另一个用途是检验两个变量的**独立性**，通过这个方法可以检查变量之间是否存在某种关联。

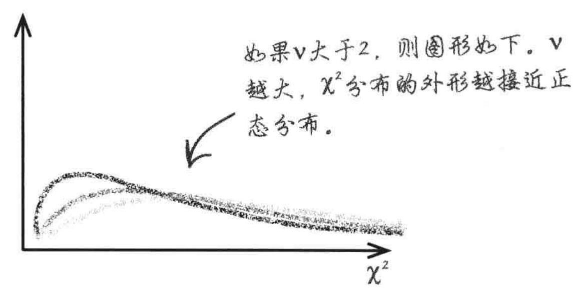
X^2分布用到一个参数——希腊字母ν，读作“纽”，让我们看看ν如何影响概率分布的形状。

当ν等于1或2



当ν等于1或2时，X^2分布为一条先高后低的平滑曲线，其形状像一个倒立的J。检验统计量等于较小数值的概率远远高于等于较大数值的概率，这就是说，观察频数有可能接近期望频数。

当ν大于2



当ν大于2时，X^2分布的形状发生改变——随着X^2递增，图形先低，后高，再低，其外形沿着正向扭曲，但当ν很大时，图形接近正态分布。

若你正在使用具有特定参数ν的X^2分布以及检验统计量X^2，可简单记作：

$$X^2 \sim \chi^2(\nu)$$

X^2符合X^2分布，给定值为ν。

看上去像X，但更显扭曲。

ν表示自由度

前面讲到ν如何影响χ^2分布的形状，如何求出ν呢？

ν为**自由度**数目，即用于计算检验统计量χ^2的独立变量的数目，或可以说是独立信息段的数目。让我们结合实际进行说明。

下面回顾一下老虎机的观察频数和期望频数：

x	观察频数	期望频数
-2	965	977
23	10	8
48	9	8
73	9	6
98	7	1

自由度数目等于我们要计算的期望频数的数目——计算时要考虑我们所受到的各种限制。

为了计算检验统计量X^2，我们必须计算所有的期望频率，也就是必须计算5个期望频数。进行计算时要记住一点：期望频数总和与观察频数总和必须相同——这就是说，我们进行计算时受到1个限制。

那么ν是多少？

为了算出ν，我们取所计算过的信息的数目，减去所受到的限制的数目。为了算出检验统计量X^2，我们必须计算5个独立信息，同时受到1个限制。于是，自由度的计算结果为：

$$
\begin{aligned}
\nu &= 5 - 1 \\
&= 4
\end{aligned}
$$

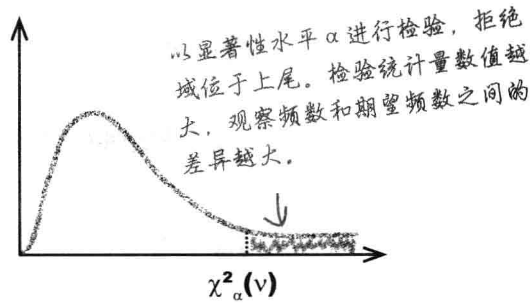
以上结果还可以这样理解：我们必须利用概率分布计算4个期望频数；至于最后一个频数，则可以先求出总期望频数，再求出最后一个频数。

一般说来，

$$\nu = (\text{组数}) - (\text{限制数})$$

显著性是多少？

我们如何利用χ^2分布指出观察频数和期望频数之间的差异显著性？和其他假设检验一样，这都取决于显著性水平。



用χ^2分布进行的检验为单尾检验，右尾被作为拒绝域。于是，通过查看检验统计量是否位于右尾的拒绝域以内，你就可以判定根据期望分布得出的结果的可能性。

如果用显著性水平α进行检验，则可以写作：

$$\chi^2_\alpha(\nu)$$

那么如何求χ^2分布的拒绝域呢？我们可以使用χ^2概率表。

如何使用χ^2概率表

为了求出临界值，首先应找出自由度ν以及显著性水平α。在第一列查找ν，第一行查找α，交点即x值，从$P(\chi^2\alpha(\nu) \geq x) = \alpha$得出临界值。

例如，以5%为显著性水平，8为自由度进行检验，若要求临界值，则在第一列查找8，第一行查找0.05，查出数值15.51。因此，只要检验统计量X^2大于15.51，则在显著性水平为5%、自由度为8的情况下，检验统计量就位于拒绝域以内。

这一列为0.05。

	尾部概率α									
ν	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001
1	1.32	1.64	2.07	2.71	[illegible]	5.02	5.41	6.63	7.88	9.14	10.83
2	2.77	3.22	3.79	4.61	[illegible]	7.38	7.82	9.21	10.60	11.98	13.82
3	4.11	4.64	5.32	6.25	[illegible]	9.35	9.84	11.34	12.84	14.32	16.27
4	5.39	5.99	6.74	7.78	[illegible]	11.14	11.67	13.28	14.86	16.42	18.47
5	6.63	7.29	8.12	9.24	[illegible]	12.83	13.39	15.09	16.75	18.39	20.51
6	7.84	8.56	9.45	10.64	[illegible]	14.45	15.03	16.81	18.55	20.25	22.46
7	9.04	9.80	10.75	12.02	[illegible]	16.01	16.62	18.48	20.28	22.04	24.32
8	[illegible]	[illegible]	[illegible]	[illegible]	15.51	17.53	18.17	20.09	21.95	23.77	26.12
9	11.39	12.24	13.29	14.68	[illegible]	19.02	19.68	21.67	23.59	25.46	27.88

这一行为ν = 8。

这是8和0.05的交点。

χ^2假设检验

下面是用χ^2分布进行假设检验的几大步骤：

1. **确定要进行检验的假设及其备择假设**
2. **求出期望频数和自由度**
3. **确定用于做决策的拒绝域**
4. **计算检验统计量X^2**
5. **查看检验统计量是否位于拒绝域以内**
6. **作出决策**

这些步骤和前面的步骤一样

这些步骤和前面提到的步骤不同

看着眼熟吗？大部分步骤都和其他假设检验完全一样，也就是说，这个过程与前面讲过的过程完全相同。

世上没有傻问题

问： 这么说χ^2检验其实就是假设检验的特殊形式？

答： 是的，正是如此。检验步骤完全和前文讲过的步骤一样。

问： 检验时总是使用右尾吗？

答： 是的，假设检验总是使用右尾。这是因为检验统计量越大，观察频数与期望频数的差别越大。

问： 我想我在前面看到过自由度这个术语，对不对？

答： 没错，前面看到过。还记得我们讲过如何用t分布建立置信区间吗？对，t分布也用到了自由度。

问： 我想以前是把自由度叫做df的，而不是ν，我记错了吗？

答： 一点儿没有错。不同课本有不同的约定，我们用的是ν。反正，它们意思相同。

问： 我想在网上查找χ^2分布的信息。该怎么查找呢？要输入希腊字母吗？

答： 查找“卡方”即可。χ^2也写作“卡方”。

练习

你的任务是，在5%的显著性水平下，看看是否有足够的证据判定老虎机被人动了手脚。请按所给步骤进行计算。

1. 要检验的原假设是什么？备择假设是什么？

2. 自由度为4，5%水平的拒绝域是多少？

3. 检验统计量是多少？ 提示：前面已经计算过。

4. 检验统计量是在拒绝域以内还是在拒绝域以外？

5. 你将接受还是拒绝原假设？



你的任务是，在5%的显著性水平下，看看是否有足够的证据判定老虎机被人动了手脚。请按所给步骤进行计算。

1. 要检验的原假设是什么？备择假设是什么？

H_0：老虎机每局收益符合如下概率分布。

x	−2	23	48	73	98
P(X = x)	0.977	0.008	0.008	0.006	0.001

H_1：老虎机每局收益不符合以上概率分布。

2. 自由度为4，5%水平的拒绝域是多少？

从概率表上查得$\chi^2_{5\%}(4) = 9.49$，即拒绝域为$X^2 > 9.49$的范围。

3. 检验统计量是多少？

检验统计量为X^2，前面已经计算过，为38.272。

4. 检验统计量是在拒绝域以内还是在拒绝域以外？

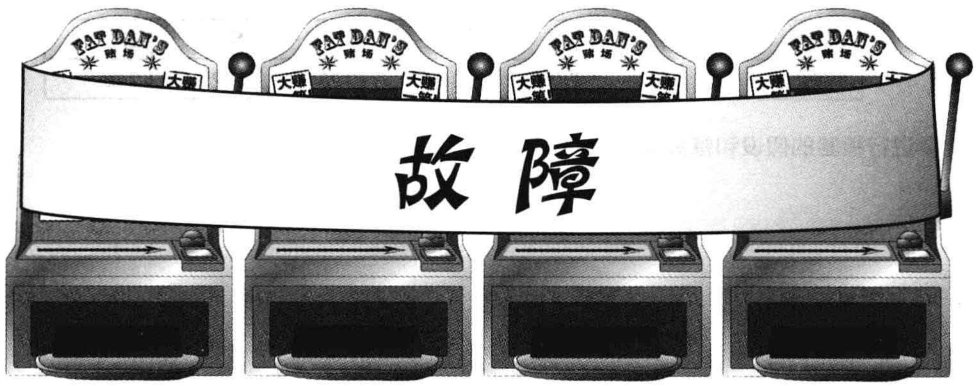
X^2的数值为38.27，且由于拒绝域为$X^2 > 9.49$，因此X^2位于拒绝域以内。

5. 你将接受还是拒绝原假设？

χ^2的数值位于拒绝域以内，于是我们拒绝原假设。即，我们有充足的证据拒绝上述“老虎机每局收益符合如下概率分布”这个原假设。

你解开了老虎机之谜

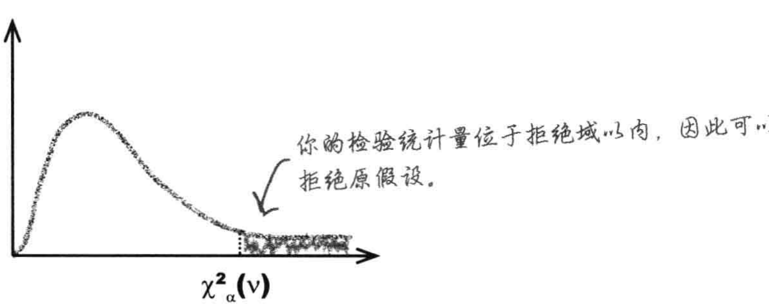
通过小心运用χ^2分布，你发现有充足的证据证明老虎机不符合赌场期望它们符合的概率分布。肥蛋十分感激你，是你的证据说明老虎机被人动了手脚。他把这些老虎机停了，免得赔钱。



让我们总结一下你的解答步骤。

首先，你得到了老虎机的一组观察频数，然后假定这些频数符合某种特定的概率分布并算出了期望频数。然后你算出自由度和检验统计量χ^2，通过χ^2可以看出观察频数和期望频数之间的总偏差。

然后，你从χ^2概率表查出显著性为5%时的拒绝域，经过与检验统计量进行比较，你发现有足够的证据判定：老虎机被人动过手脚，所以才会多赔钱。



这种假设检验称为**拟合优度**检验——它检验观察频数是否和假设的概率分布相吻合。若你有一组数据，并希望这组数据符合某种分布，为了看看这组数据是否确实符合这种分布，则可以用拟合优度检验。



加强练习

肥蛋认为骰子有问题。下表中列出了一个骰子的观察频数，查看这些数据，并以1%的显著性水平进行检验，看看是否有足够的证据说明的确存在不公正。请按照我们给出的步骤进行。

下面是观察频数：

数值	**1**	**2**	**3**	**4**	**5**	**6**
频数	107	198	192	125	132	248

第1步：决定要进行检验的假设和备择假设。

第2步：求期望频数和自由度。

首先填写骰子的期望频数，应考虑掷骰子的总次数以及每个数值的掷出概率。X代表掷出的骰子点数。

x	**观察频数**	**期望频数**
1	107	
2	198	
3	192	
4	125	
5	132	
6	248	

求出期望频数后，再算算自由度是多少？

自由度计算方法和老虎机用的方法相同。

第3步：确定用于做决策的拒绝域。

将会用到显著性水平和自由度。

第4步：计算检验统计量χ^2。

可以用第2步算出的观察频数和期望频数进行计算。

第5步：看看检验统计量是否位于拒绝域以内。

第6步：作出决策。

加强练习解答

肥蛋认为骰子有问题。下表中列出了一个骰子的观察频数，查看这些数据，并以1%的显著性水平进行检验，看看是否有足够的证据说明的确存在不公正。请按照我们给出的步骤进行。

下面是观察频数：

数值	1	2	3	4	5	6
频数	107	198	192	125	132	248

第1步：决定要进行检验的假设和备择假设。

为了检验骰子是否公正，我们必须确定是否有足够证据说明骰子不公正。
于是：

H_0：骰子公正，每一面数值的掷出几率都相同，即每一面数值的发生概率为1/6。

H_1：骰子不公正。

第2步：求期望频数和自由度。

首先填写骰子的期望频数，应考虑掷骰子的总次数以及每个数值的掷出概率。X代表掷出的骰子点数。

x	观察频数	期望频数
1	107	167
2	198	167
3	192	167
4	125	167
5	132	167
6	248	167

总期望频数必须与总观察频数相符。如果将所有观察频数相加，结果为1002。

每一面数值的掷出概率为1/6，因此每一面数值的期望频数为1002/6 = 167。

求出期望频数后，再算算自由度是多少？

我们必须求出6个期望频数，其总和等于1002。即我们必须求出6个信息，同时受到1个限制。因此：

$$\nu = 6 - 1$$
$$= 5$$

第3步：确定用于做决策的拒绝域。

将会用到显著性水平和自由度

从概率表查出$\chi^2_{1\%}(5) = 15.09$，于是拒绝域为$\chi^2 > 15.09$的范围。

第4步：计算检验统计量χ^2。

可以用第2步算出的观察频数和期望频数进行计算。

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

$= (107-167)^2/167 + (198-167)^2/167 + (192-167)^2/167 + (125-167)^2/167 + (132-167)^2/167 + (248-167)^2/167$

$= (-60)^2/167 + (31)^2/167 + (25)^2/167 + (-42)^2/167 + (-35)^2/167 + (81)^2/167$

$= (3600 + 961 + 625 + 1764 + 1225 + 6561)/167$

$= 14736/167$

$= 88.24$

第5步：看看检验统计量是否位于拒绝域以内。

拒绝域由$\chi^2 > 15.09$决定，由于$\chi^2 = 88.24$，因此检验统计量位于拒绝域内。

第6步：作出决策

由于你的检验统计量位于拒绝域内，说明在显著性水平为1%的情况下，有足够的证据拒绝原假设，于是你接受备择假设：骰子不公正。



χ^2拟合优度检验对相当多的概率分布都有效。

只要你得到一组观察频数，且能算出期望频数，就可以用χ^2分布检验任何概率分布的拟合优度。

最大的困难在于自由度ν的计算，下面是最常用的一些概率分布的自由度，可在进行χ^2拟合优度检验时使用。

p是成功概率，或者说是总体的成功概率。

分布	条件	ν
二项分布	已知p	$\nu = n - 1$
	未知p，必须通过观察频数进行估计	$\nu = n - 2$
泊松分布	已知λ	$\nu = n - 1$
	未知λ，必须通过观察频数进行估计	$\nu = n - 2$
正态分布	已知μ和σ^2	$\nu = n - 1$
	未知μ和σ^2，必须通过观察频数进行估计	$\nu = n - 3$

n是观察频数总数。

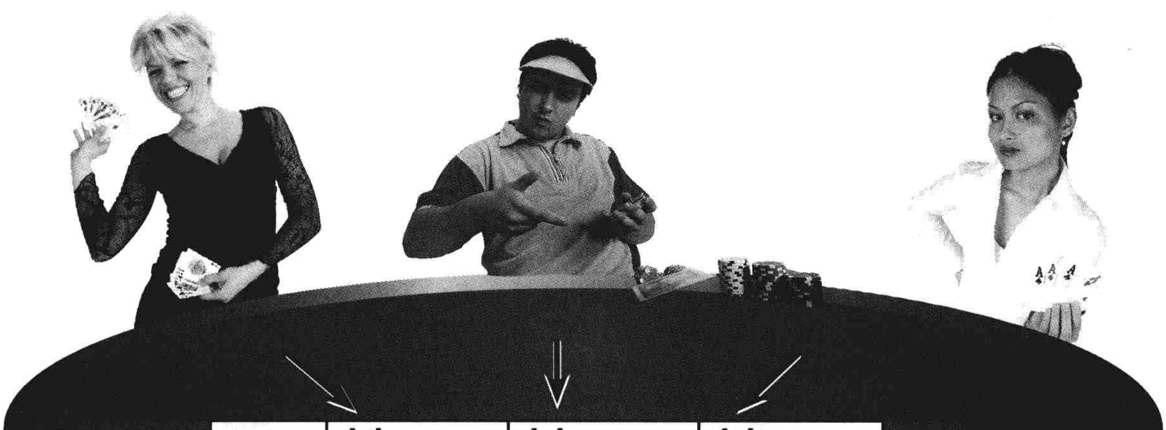
λ是一定区间内的发生率。

肥蛋遇到了新问题

前面你调查了老虎机是否被人动过手脚，用的是拟合优度检验，目的是判断观察频数是否与所期望的概率分布相吻合。肥蛋还有一个问题，这一次事关内部员工。

肥蛋觉得有一位负责二十一点赌桌的庄家赔付的钱高于合理值。你能判断一下是否有显著证据证明肥蛋的怀疑没错？

下面是负责赌桌的三位庄家：



	庄家A	庄家B	庄家C
赢	43	49	22
平	8	2	5
赔	47	44	30

这是每一局赌局的结果。

这是每一位庄家的观察结果。

我们需要找到某种方法，检验赌局结果是否取决于坐庄赌局的庄家。

动动脑

检验以上假设需要知道哪些条件？

χ^2 分布可以检验独立性

前面讲到χ^2分布可用于进行拟合优度检验，χ^2分布的用途不仅如此，它还能用于进行**独立性**检验。

独立性χ^2检验可用于判断两种因素是否相互独立，或两者是否看上去互有联系。这正合我们对庄家的检验要求 —— 我们要检验在二十一点赌局中坐庄的庄家是否对赌局输赢有影响。换句话说，我们假定庄家的选择与输赢无关 —— 除非有足够的证据可以反驳这一点。

独立性检验的过程与拟合优度的检验过程相同：设立一个假设，用观察频数和期望频数计算X^2检验统计量，然后查看结果是否落在拒绝域以内。



为了计算检验统计量X^2，我们需要知道期望频数。

这说明我们需要通过观察频数算出期望频数，这得依靠概率……

可用概率求出期望频数

期望概率可通过几个步骤求得。

首先，算出赌局结果和庄家总频数以及各项的总和，例如可列出下表，这叫做**列联表**。

	庄家A	庄家B	庄家C	合计
赢	43	49	22	114
平	8	2	5	15
赔	47	44	30	121
合计	98	95	57	250

赢局次数

庄家A合计

总和

现在我们可以用以上信息求出每一位庄家的赢局期望频数。

让我们先求出庄家A的赢局期望频数。

首先，我们可以用以上总和求出得到一个特定结果的概率，或者求出某位庄家的概率。例如，为了求出赢局概率，可以用赢局合计除以总和：

$$P(\text{赢}) = \frac{\text{赢局合计}}{\text{总和}}$$

同样，可用庄家A的坐庄次数除以总和，求出庄家A的坐庄概率：

$$P(A) = \frac{\text{合计A}}{\text{总和}}$$

现在，按照我们的假设，如果庄家和赌局结果相互独立，那么，通过将两种概率相乘，可以求出庄家A坐庄时出现赢局的概率，即：

$$P(A\text{庄赢局}) = \frac{\text{赢局合计}}{\text{总和}} \times \frac{\text{A合计}}{\text{总和}}$$

第4章讲过，对于独立事件：$P(A \cap B) = P(A) \times P(B)$。

动动脑

我们如何利用以上公式求出庄家A的赢局期望频数？

频数是多少？

前面求出了庄家A的赢局概率，我们希望通过这个结果求出赢局的期望频数。为此只要将庄家A的赢局概率乘以总和即可，于是：

$$期望频数 = \cancel{总和} \times \frac{赢局合计}{\cancel{总和}} \times \frac{A总计}{总和}$$

$$= \frac{赢局合计 \times A总计}{总和}$$

即，为了求出庄家A的赢局期望频数，可用所有赢局合计数目乘以庄家A的赌局数目，然后除以总和。

一般我们如何求频数？

将以上结果推而广之，可以得到一个求频数的通用公式：为了求出特定行和特定列形成的组合的期望频数，可用每行合计乘以每列合计，然后除以总和。

$$\textbf{期望频数} = \frac{\textbf{行合计} \times \textbf{列合计}}{\textbf{总和}}$$

求出所有期望频数后，即可用它计算出检验统计量X^2—— 这与前面的检验统计量相同。因此需要计算：

$$X^2 = \sum \frac{(O - E)^2}{E}$$

用每一个观察频数减去期望频数，所得结果求平方，再除以期望频数，最后加总。

关键是：务必将每一个观察频数和每一个相应的期望频数都计算在内。



练习

下表显示了各位庄家的观察频数。你的任务是算出所有期望频数。

这些是观察频数。→

	庄家A	庄家B	庄家C	合计
赢	43	49	22	114
平	8	2	5	15
赔	47	44	30	121
合计	98	95	57	250

(行合计 × 列合计)/总和

在这张表里填入每个期望频数。→

	庄家A	庄家B	庄家C
赢	(114×98)/250=44.688		
平	(15×98)/250=5.88		
赔	(121×98)/250=47.432		

求出所有期望频数后，计算检验统计量X^2。下表可以提供帮助：第一列给出了所有观察频数，第二列是相应的期望频数，只要将第三列的所有数字加起来，就可以得到检验统计量。

用前两列的数值可以帮助你完成计算。

	观察	期望	$\frac{(O-E)^2}{E}$
A	43	44.688	$(43-44.688)^2/44.688 = 2.85/44.688 = 0.064$
	8	5.88	$(8-5.88)^2/5.88 = 4.4944/5.88 = 0.764$
	47	47.432	$(47-47.432)^2/47.432 = 0.187/47.432 = 0.004$
B	49		
	2		
	44		
C	22		
	5		
	30		
	ΣO = 250	**ΣE =**	$\sum \frac{(O-E)^2}{E} =$



练习解答

下表显示了各位庄家的观察频数。你的任务是算出所有期望频数。

观察频数

	庄家A	庄家B	庄家C	合计
赢	43	49	22	114
平	8	2	5	15
赔	47	44	30	121
合计	98	95	57	250

期望频数

	庄家A	庄家B	庄家C
赢	(114×98)/250=44.688	(114×95)/250=43.32	(114×57)/250=25.992
平	(15×98)/250=5.88	(15×95)/250=5.7	(15×57)/250=3.42
赔	(121×98)/250=47.432	(121×95)/250=45.98	(121×57)/250=27.588

求出所有期望频数后，计算检验统计量X^2。下表可以提供帮助：第一列给出了所有观察频数，第二列是相应的期望频数，只要将第三列的所有数字加起来，就可以得到检验统计量。

	观察	期望	$\frac{(O-E)^2}{E}$
A	43	44.688	$(43-44.688)^2/44.688 = 2.85/44.688 = 0.064$
A	8	5.88	$(8-5.88)^2/5.88 = 4.4944/5.88 = 0.764$
A	47	47.432	$(47-47.432)^2/47.432 = 0.187/47.432 = 0.004$
B	49	43.32	$(49-43.32)^2/43.32 = 5.68/43.32 = 0.131$
B	2	5.7	$(2-5.7)^2/5.7 = 13.69/5.7 = 2.402$
B	44	45.98	$(44-45.98)^2/45.98 = 3.9204/45.98 = 0.085$
C	22	25.992	$(22-25.992)^2/25.992 = 15.936/25.992 = 0.613$
C	5	3.42	$(5-3.42)^2/3.42 = 2.4964/3.42 = 0.730$
C	30	27.588	$(30-27.588)^2/27.588 = 5.817/27.588 = 0.211$
	$\Sigma O = 250$	$\Sigma E = 250$	$\sum \frac{(O-E)^2}{E} = 5.004$

这是你的检验统计量

我们还需要计算自由度

为了用χ^2分布求观察频数的显著性，还需求出最后一个值：ν，即自由度值。

前面讲过，自由度是在考虑限制条件的情况下，可以自由选择的独立信息的数目。这说明我们要查看有多少个需要独立计算的期望频数，再减去限制条件数目。

首先，让我们求要计算的期望频数的总数目。我们必须算出三位庄家的期望频数以及三种可能结果，于是期望频数为 $3 \times 3 = 9$ 。

我们必须求出 $3 \times 3 = 9$ 种期望频数。

	庄家A	庄家B	庄家C
赢			
平			
赔			

对于每一行每一列，我们实际上只需要计算两个期望频数。我们已经知道总频数是多少，因此可以选择第三个频数，使得所有频数相加等于正确结果。也就是说，我们其实只需要计算其中4个期望频数，其余5个频数可以根据已知的总频数进行推导。

我们只需要计算这几个频数，其余的则可以借助每一行和每一列的总频数求出。

	庄家A	庄家B	庄家C
赢			
平			
赔			

利用合计可以求出最后一行和最后一列的结果。

由于必须算出4个期望频数，于是自由度就等于这个数目——共需要计算4个独立信息；算出这些频数后，其余频数自然就知道了。即：$\nu = 4$。

另一种得知自由度的方法是：我们总共需要计算9个数值，其中5个不用独立进行计算。用前面的公式可计算$\nu = 9 - 5 = 4$。

加强练习

以1%的显著性水平进行假设检验，看看赌局结果是否独立于坐庄的庄家。下面给出步骤提示，不过别忘了，有些结果前面已经算出来了。

1. 确定要进行检验的假设及其备择假设。

2. 求出期望频数和自由度。

3. 确定用于做决策的拒绝域。

4. 计算检验统计量X^2。

5. 看看检验统计量是否位于拒绝域以内。

6. 作出决策。

纸张足够，请尽情计算。



加强练习解答

以1%的显著性水平进行假设检验，看看赌局结果是否独立于坐庄的庄家。下面给出步骤提示，不过别忘了，有些结果前面已经算出来了。

1. 确定要进行检验的假设及其备择假设。

2. 求出期望频数和自由度。

3. 确定用于做决策的拒绝域。

4. 计算检验统计量 X^2。

5. 看看检验统计量是否位于拒绝域以内。

6. 作出决策。

第1步：

我们要检验赌局输赢结果是否独立于坐庄的庄家，于是：

H_0：赌局输赢结果和坐庄的庄家没有关系。

H_1：赌局输赢结果和坐庄的庄家有关系。

第2步：

我们在590页求出了期望频数，并得出自由度为4。

第3步：

从概率表查出 $\chi^2_{1\%}(4) = 13.28$，因此拒绝域由 $X^2 > 13.28$ 决定。

第4步：

在590页我们还用期望频数算出了 $X^2 = 5.004$。

第5步：

拒绝域由 $X^2 > 13.28$ 给出，因此 X^2 位于拒绝域以外。

第6步：

由于 X^2 位于拒绝域以外，因此我们接受原假设：没有足够的证据证明赌局结果和庄家之间有关系。

世上没有傻问题

问： 我还是不太确定自己是否理解了庄家自由度的算法。为什么有4个自由度？

答： 自由度是这样计算的：查看需要计算几个期望频数，然后再看这些频数中有几个能够仅仅通过观察每一列和每一行的观察频数合计即可得出。

问题中包含三名庄家，三组结果，如果用列联表进行计算，则各列和各行的期望频数合计必须等于观察频数合计。这说明，只要算出任意行或任意列的前2个频数，就可以通过合计求出最后一个频数。因此，完全自行进行计算的频数只有2×2，因此自由度为4。

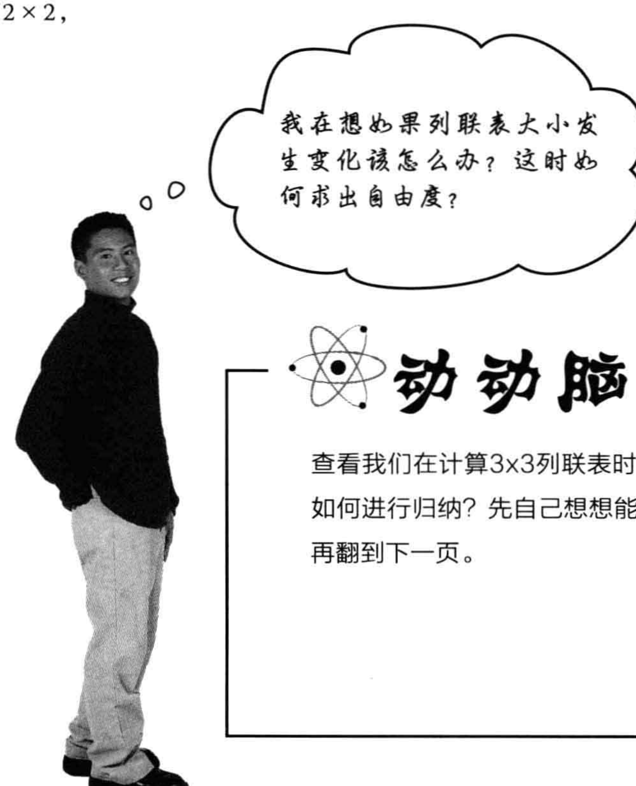
问： 除了拟合优度检验和独立性检验，χ^2分布还有其他用途吗？

答： χ^2分布主要就是这两种用途，记住，你几乎可以用它检验任意概率分布的拟合优度。例如，可以检验观察频数是否符合特定二项分布。

问： 我应该以任意显著性水平进行检验吗？

答： 看情况。与其他假设检验一样，显著性水平越小，为了拒绝原假设所需要的证据越强。

检验时常用的显著性水平为5%和1%。



动动脑

查看我们在计算3x3列联表时的做法，你觉得可以如何进行归纳？先自己想想能不能找到办法，然后再翻到下一页。

自由度计算方法归纳

前面讲到3×3列联表的自由度计算，如何归纳这个算法呢？

假设你正在对两个变量进行比较，且一个变量有h行，另一个变量有k列，行和列的合计有办法知道。假设要求自由度的数目。

	列1	…	列k-1	列k
行1				
…				
行h-1				
行h				

每一行都对应着k列。你有办法知道每一行的合计，因此实际上只要算出（k-1）列就行了，由于该行的总频数已知，因此第k列自然就会知道。

	列1	…	列k-1	列k
行1				

利用该行合计，可以求出第k列。

这些是需要计算的

列的计算与此相似。每一列都对应h行，你有办法知道每一列的合计，因此可以算出（h-1）行，由于该列的总频数已知，因此第h行自然就会知道。

	列1
行1	
…	
行h-1	
行h	

你需要计算h-1行的频数。

你可以用列合计算出第h行。

得出算式……

综合以上结果，需要计算的期望频数的总数目为(k-1)×(h-1)，即，如果有一张大小为h × k的表格，就可以通过下列算式得出自由度：

$$\nu = (h - 1) \times (k - 1)$$

	列1	…	列k-1	列k
行1				
…				
行h-1				
行h				

必须计算(h-1) × (k-1)个期望频数，因此自由度为(h-1) × (k-1)。

动动笔

肥蛋又招聘了两名庄家。现在自由度是多少？赌局结果保持不变。

动动笔解答

肥蛋又招聘了两名庄家。现在自由度是多少？赌局结果保持不变。

由于肥蛋又招聘了两名庄家，因此列联表变为3×5。

A、B、C是原来的庄家，肥蛋又招聘了两名庄家。

	庄家A	庄家B	庄家C	庄家D	庄家E
赢					
平					
赔					

自由度算式为(h-1)×(k-1)，其中h为行数，k为列数，于是：

$$\nu = 2 \times 4$$
$$= 8$$

要点

- 通过X^2分布可以进行拟合优度检验和变量独立性检验。
- 检验统计量为

$$X^2 = \sum \frac{(O-E)^2}{E}$$

其中O指的是观察频数，E指的是期望频数。

- 如果在X^2分布中用X^2作为检验统计量，则写作：

$$X^2 \sim \chi^2_{\alpha}(\nu)$$

其中ν为自由度，α为显著性水平。

- 在拟合优度检验中，ν等于组数减去限制数。
- 在两个变量的独立性检验中，若列联表为h行k列，则：

$$\nu = (h-1) \times (k-1)$$

你救了肥蛋赌场

多亏你精通χ^2分布，你刨根究底地调查被动过手脚的赌具，对实际结果和期望结果之间的可解释差异进行了辨析，还以一定显著性水平研究了可疑的行为。

你的工作让肥蛋开心起来。谢谢！肥蛋现在知道哪个赌博游戏需要调查，而庄家们则保住了自己的饭碗。下一次来这里的时候要通知肥蛋哦，他将多给你一些筹码 —— 分文不取。

干得漂亮！



肥蛋承诺免费给你一堆赌场筹码！



加强练习

肥蛋觉得有一个或多个庄家在控制轮盘赌的结果，下面是关于每一位庄家的停球颜色的观察频数数据。请以5%的置信度进行检验，看看球位颜色是否与庄家相互独立，或者说，是否有足够证据证明可能存在隐情。

	庄家A	庄家B	庄家C
红	375	367	357
黑	379	336	362
绿	46	37	41

第1步：决定要进行检验的假设及其备择假设。

第2步：使用下列期望频数表，求期望频数和自由度。

提示：首先填写各行、各列的合计值，这些合计值与前面的观察频数合计值是相同的。

	庄家A	庄家B	庄家C	合计
红	1099×800/2300=382.3	1099×740/2300=353.6		
黑	1077×800/2300=374.6			
绿	124×800/2300=43.1			
合计	800			

第3步：确定用于决策的拒绝域。

第4步：利用下表，计算检验统计量 X^2。

	观察	期望	$\frac{(O-E)^2}{E}$
	375	382.3	$(375-382.3)^2/382.3 = 53.29/382.3 = 0.139$
A	379	374.6	$(379-374.6)^2/374.6 = 19.36/374.6 = 0.005$
	46	43.1	$(46-43.1)^2/43.1 = 8.41/43.1 = 0.195$
	367	353.6	$(367-353.6)^2/353.6 = 179.56/353.6 = 0.508$
B	336		
	37		
	357		
C	362		
	41		
	$\Sigma O =$	$\Sigma E =$	$\sum \frac{(O-E)^2}{E} =$

第5步：查看检验统计量是否位于拒绝域以内。

第6步：作出决策。

加强练习解答

肥蛋觉得有一个或多个庄家在控制轮盘赌的结果，下面是关于每一位庄家的停球颜色的观察频数数据。请以5%的置信度进行检验，看看球位颜色是否与庄家相互独立，或者说，是否有足够证据证明可能存在隐情。

	庄家A	庄家B	庄家C
红	375	367	357
黑	379	336	362
绿	46	37	41

第1步：决定要进行检验的假设及其备择假设。

你要检验球位颜色是否与庄家相互独立，因此：

H_0：轮盘球位颜色与庄家相互独立。

H_1：球位颜色与庄家相互不独立

第2步：使用下列期望频数表，求期望频数和自由度。

将每一行与每一列的合计相乘，再除以总和，得出期望频数。

	庄家A	庄家B	庄家C	合计
红	1099×800/2300=382.3	1099×740/2300=353.6	1099×760/2300=363.1	1099
黑	1077×800/2300=374.6	1077×740/2300=346.5	1077×760/2300=355.9	1077
绿	124×800/2300=43.1	124×740/2300=39.9	124×760/2300=41.0	124
合计	800	740	760	2300

共有3行3列，用(行数 − 1)乘以(列数 − 1)，得到自由度：

$$\nu = 2 \times 2$$
$$= 4$$

第3步：确定用于决策的拒绝域。

从概率表查得$\chi^2_{5\%}(4) = 9.49$，于是拒绝域由$\chi^2 > 9.49$决定。

第4步：利用下表，计算检验统计量X^2。

	观察	期望	$\frac{(O-E)^2}{E}$
A	375	382.3	$(375-382.3)^2/382.3 = 53.29/382.3 = 0.139$
	379	374.6	$(379-374.6)^2/374.6 = 19.36/374.6 = 0.005$
	46	43.1	$(46-43.1)^2/43.1 = 8.41/43.1 = 0.195$
B	367	353.6	$(367-353.6)^2/353.6 = 179.56/353.6 = 0.508$
	336	346.5	$(336-346.5)^2/346.5 = 110.25/346.5 = 0.318$
	37	39.9	$(37-39.9)^2/39.9 = 8.41/39.9 = 0.211$
C	357	363.1	$(357-363.1)^2/363.1 = 37.21/363.1 = 0.102$
	362	355.9	$(362-355.9)^2/355.9 = 37.21/355.9 = 0.105$
	41	41.0	$(41-41)^2/41 = 0/41 = 0$
	$\Sigma O = 2300$	$\Sigma E = 2300$	$\sum\frac{(O-E)^2}{E} = 1.583$

这表示检验统计量为$X^2 = 1.583$。

第5步：查看检验统计量是否位于拒绝域以内。

拒绝域由$X^2 > 9.48$给定，由于$X^2 = 1.583$，因此检验统计量位于拒绝域以外。

第6步：作出决策。

由于检验统计量位于拒绝域以外，因此在显著性水平为5%的情况下，没有充足的理由可以拒绝原假设。即，接受原假设：球位颜色和庄家相互独立。

15 相关与回归

我的线条如何？



你是否曾经为某两件事的相互关系困惑不已？

前面讲过的统计量只描述一个变量——如个人身高、篮球队员得分或是糖球口味持续时间，但是，另外还有一些统计量可以说明**变量之间的关系**。了解事物的相互关系可以丰富你的信息，让你了解真相，使你立于不败之地。来吧，让我们为你介绍**发现事物关系的秘诀**：相关与回归。

永远不要相信天气

露天音乐会是最棒的音乐会——起码这两位帅哥是这么想的，他们承接组织一场商业性露天音乐会，夏季的票房看来有大卖的希望。

今天的音乐会有望成为演出以来的最佳场次,乐队已经开始练习。只是，天边飘来一片乌云……



不消片刻，天色阴沉下来，气温骤降，雨似乎要下起来了。更糟糕的是，票房受创，小伙子们麻烦了，再出这种事他们可赔不起。

小伙子们希望自己能够根据预计天晴时数（小时）预测出音乐会听众人数。这样一来，他们就可以衡量阴天可能给听众人数造成的影响。如果听众人数将少于3,500人——这时票房收入将无法抵消成本费用，那么他们就取消音乐会。

他们需要**你**帮帮忙。

让我们分析天晴时数和听众人数

下面是样本数据，给出了不同场次的预计天晴时数和音乐会听众人数的关系数据。利用这些数据，我们如何基于当天预计天晴时数（小时）估计出票情况？

天晴时数（小时）	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
音乐会听众人数（百人）	22	33	30	42	38	49	42	55

这简单。我们可以求均值、标准差，再观察分布，那样就全都清楚了。



大多数时候，我们只需要如此这般行事就能预测各种可能结果。

这一次的问题在于，我们该求哪些数据的均值和标准差？我们该以音乐会听众人数作为计算基础，还是该以天晴时数作为计算基础？二者都没有给出我们所需要的全部信息——我们不能只使用一组数据，而是两组数据都要使用。

前面我们只讲过独立随机变量，相关变量还没有讲到。我们可以假设，如果天气不好，则露天音乐会出现高上座率的概率将比天气好时的概率低。可是我们如何为这种关系建立模型呢？我们如何利用这个模型按照天晴时数预测听众上座率呢？

这取决于数据类型。

动动脑

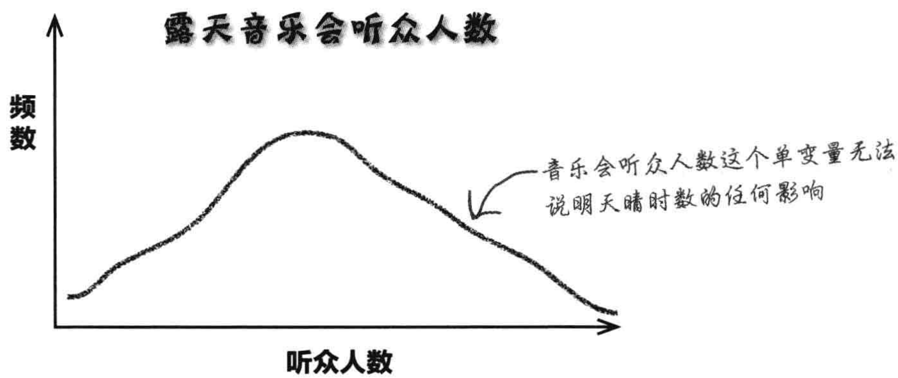
你会如何建立模型描述两组数据的关系？

数据类型探讨

迄今为止，我们所使用的变量都是单变量。

单变量数据考虑的是一个单一变量的频数或概率，例如，单变量数据可以描述赌场收益或是统计邦新娘的体重，在这两种情况下，所描述的对象各只有一种。

单变量数据无法显示多组数据之间的关系，例如，如果用一个单变量数据描述一场露天音乐会的听众人数，那么这个变量无法说明当天预计天晴时数的任何情况，而只能给出音乐会听众人数。



所以，如果我们需要了解不同变量之间的关系，该怎么办？尽管单变量无法为我们提供这类信息，却有另一种类型的数据能够办到——**二变量数据**。

二变量数据面面观

对于每一个观察结果，二变量数据给出两个变量数值——而不是一个，例如，对于同一场音乐会，或者说对于同一个观察结果，二变量数据会同时给出预计天晴时数和音乐会听众人数，如下所示：

二变量数据为同一个观察结果提供两个变量数值。

天晴时数（小时）	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
音乐会听众人数（百人）	22	33	30	42	38	49	42	55

如果其中一个变量以某种方式受到控制，或者被用来解释另一个变量，则这个变量被称为**自变量**或**解释变量**，另一个变量则称为**因变量**或**反应变量**。在以上的例子中，我们希望用天晴时数预测听众人数，所以天晴时数是自变量，听众人数是因变量。

二变量数据可视化

像绘制单变量数据图形一样，你可以绘制二变量数据图形，借此了解数据模式。这种图不是依照频数或概率绘制数值，而是以x轴描述一个变量，以y轴描述另一个相应变量。借助这种图可以以可视方式体现两个变量之间的关系。

这种图叫做**散点图**或**散布图**，其绘制方法与其他图形的绘制方法相似。

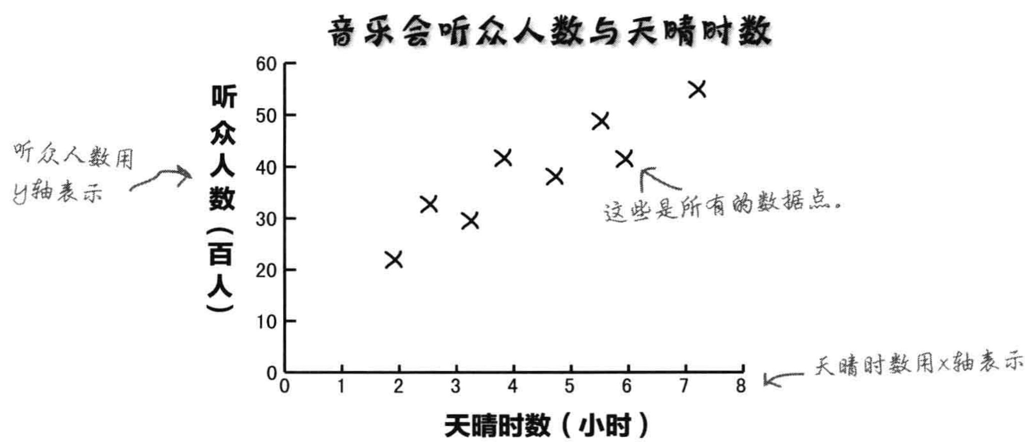
先画两条轴——横轴和纵轴，用x表示一个变量，用y表示另一个变量。自变量通常用x轴表示，因变量用y轴表示。画出坐标轴后，取每个观察结果的数值，将它们画在散点图上。

下面这张散点图显示了一场音乐会或一个观察结果中的天晴时数与音乐会听众人数的关系，由于预计天晴时数为自变量，我们将它标在x轴上，音乐会听众人数为因变量，因此用y轴表示。

天晴时数画在x轴上，听众人数画在y轴上。

x（天晴时数）	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y（听众人数）	22	33	30	42	38	49	42	55

数据在此。



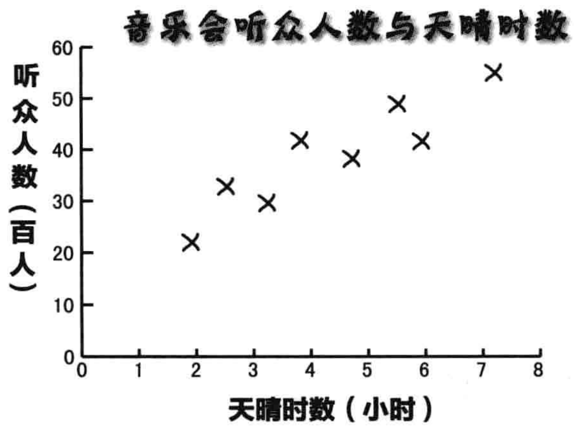
你能看出散点图如何帮助你将数据模式可视化吗?

你能看出如何借助这张图确定露天音乐会听众人数与当天预计天晴时数之间的关系吗?

动动笔

当然，我们还没有讲过如何分析二变量数据，不过让我们看看你是否能为音乐会组织者深入分析散点图。

你从图中看出了什么模式？这种模式与基础数据有何关系？如果是晴天，你对于露天音乐会听众人数有何期望？如果是阴天呢？



案件：防晒霜销量

一家防晒霜厂给了一名实习生一个任务 —— 分析防晒霜销量，看看如何以最佳方式进行品牌营销。



实习生拿到了一大堆现成的散点图，这些散点图针对防晒霜销量和各种其他因素建立了模型。厂里要求他选出这样的图形：图上的两个因素看上去存在某种关系。这对销售团队有帮助。

实习生找出的第一张图所绘制的是当天防晒霜销量与花粉量。他惊讶地发现，若花粉量高，则防晒霜销量大幅度提高。他决定告诉销售团队：他们需要考虑在广告中提到花粉量。

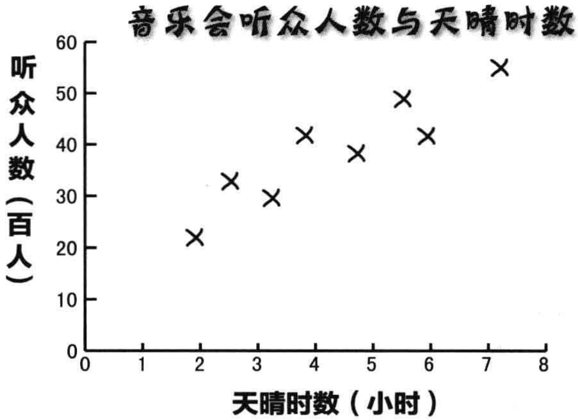
销售团队听了他的建议后，一脸茫然地看着他。你觉得销售团队应该做什么？

花粉量多会促使人们购买防晒霜吗？

动动笔解答

当然，我们还没有讲过如何分析二变量数据，不过让我们看看你是否能为音乐会组织者深入分析散点图。

你从图中看出了什么模式？这种模式与基础数据有何关系？如果是晴天，你对于露天音乐会听众人数有何期望？如果是阴天呢？



首先，从图中可以看出，数据点在图上呈直线分布，且这条线随天晴时数增加而向上爬升。看来，如果预计天晴时数相对较少，则音乐会听众人数也会减少。如果天晴时数增加，则可以期望音乐会参与人数也增加。这基本上说明，天气越晴朗，预期参加露天音乐会的人就会越多。

有一个重点需要提一下，只有在处于数据范围以内时，我们才能自信地给出这个结论，如果天晴时数小于2小时或大于7.5小时，则无数据可说明是何模式。

散点图为你指出模式

如你所见，散点图的作用在于能体现数据的实际模式，通过散点图，你可以愈发清晰地勾勒出两个变量之间的关系 —— 如果确实存在某种关系的话。

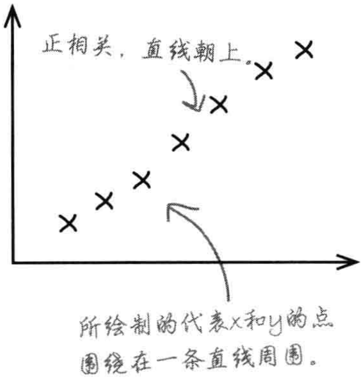
音乐会数据散点图显示出一种独特的模式 —— 数据点呈直线分布，我们将这种现象称为**相关**。

线性相关细细看

散点图显示出数据对之间的**相关性**。

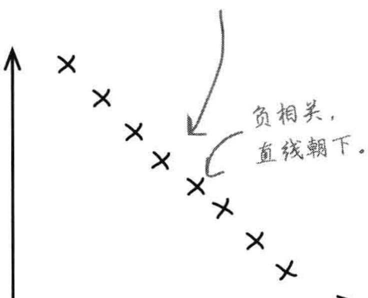
相关性即变量之间的数学关系，通过散点图上的点的独特构成模式，可以识别出散点图上的各种相关性。如果散点图上的点几乎呈直线分布，则相关性为**线性**。

让我们看看两个变量之间的相关性的几种常见类型：



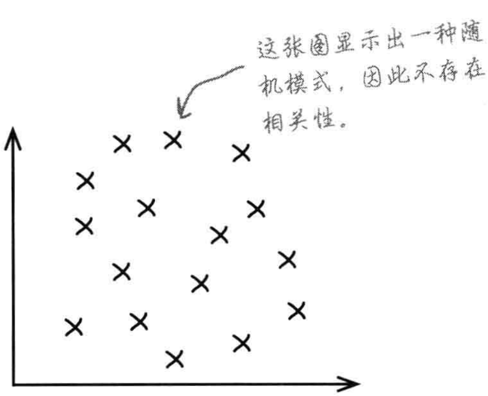
正线性相关

当x轴上的低端值对应y轴上的低端值，同时x轴上的高端值对应y轴上的高端值且呈直线分布时，为正线性相关。即随着x增长，y也呈现增长趋势。



负线性相关

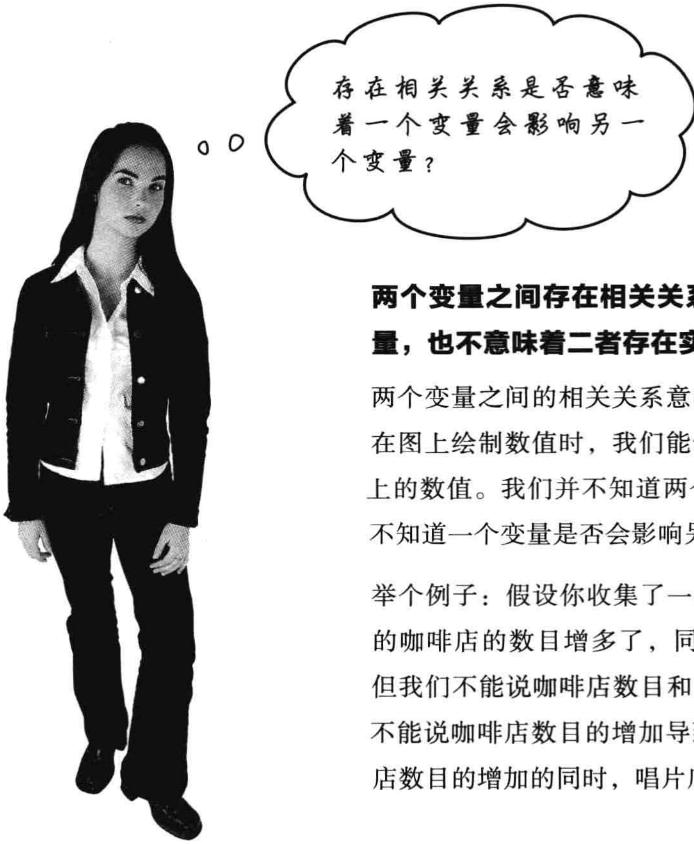
当x轴上的低端值对应y轴上的高端值，同时x轴上的高端值对应y轴上的低端值且呈直线分布时，为负线性相关。即随着x增长，y呈现下降趋势。



不相关

如果x和y的数值呈现出一种随机模式，则我们说二者不相关。

相关关系与因果关系

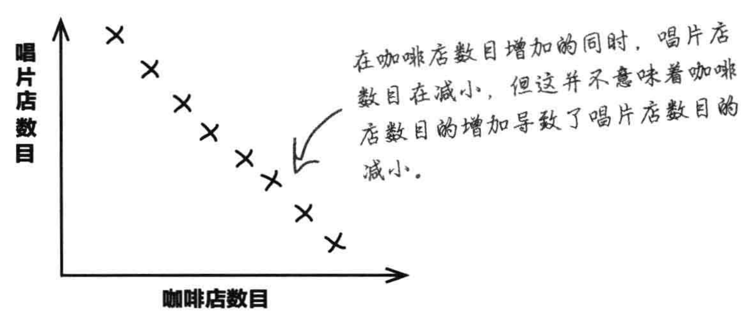


两个变量之间存在相关关系并不一定意味着一个变量会影响另一个变量，也不意味着二者存在实际关系。

两个变量之间的相关关系意味着二者之间存在某种**数学关系**，即，当我们在图上绘制数值时，我们能够看出某种模式，并能够预测出没有出现在图上的数值。我们并不知道两个变量之间是否存在**实际关系**，当然，我们也不知道一个变量是否会影响另一个变量，或是否有其他因素在发挥作用。

举个例子：假设你收集了一些数据并发现，随着时间的推移，某个小镇上的咖啡店的数目增多了，同时唱片店的数目减小了。这可能的确是实情，但我们不能说咖啡店数目和唱片店数目之间有什么实在的关系，即，我们不能说咖啡店数目的增加导致了唱片店数目的减小。我们只能说：在咖啡店数目的增加的同时，唱片店的数目减少了。

咖啡店与唱片店



破案：防晒霜销量案例

花粉量多会促使人们购买防晒霜吗？

一位销售员走到实习生身边。



“谢谢你出的主意。”她说道，“可是我们不打算用它做广告。要知道，花粉量多不会促使人们多买防晒霜。”

实习生困惑地看着她，“可散点图上不是明摆着吗，当花粉量上升时，防晒霜销量也上升。”

“确实如此。”销售员说道，“但这并不意味着花粉量多会导致销量大。在花粉量多的日子里，通常天气晴好，于是人们就会增加户外活动，人们多买防晒霜是因为他们在进行户外活动。”



世上没有傻问题

问： 这么说预计天晴时数会影响票房收入?

答： 二变量数据表明两个变量之间存在某种数学关系，但我们无法用二变量数据证明原因和结果。凭直觉，若天气晴朗，去听音乐会的人会增多，但我们不能肯定地说是天晴造成了人们去听音乐会。我们还需要做更多调查，因为可能存在其他因素。

问： 其他因素？例如?

答： 比如参加演出的艺术家的名气。如果一位著名艺术家正在举办一场音乐会，那么，无论天气如何，粉丝们都会去听音乐会。类似道理，一位冷门艺术家则不可能受到粉丝们的同样追捧。

问： 散点图用的是总体数据还是样本数据?

答： 都能用。大多数时候，你实际上是在用样本，但无论是用样本还是用总体，绘制散点图的过程都相同。

问： 如果两个变量之间有关系，必须是线性关系吗?

答： 相关性量度的是线性关系，但并不是所有关系都是线性的。例如，两个变量之件的某种强关系可能是一条特别的曲线，例如$y=x^2$。不过，我们在本章中只介绍线性关系。



我们需要预测音乐会听众人数

前面讲到什么是二变量数据，以及散点图如何体现两个变量之间是否存在数学关系，不过还没有讲过如何利用散点图进行预测。

接下来我们就需要看看，如何利用已有数据根据预计天晴时数预测音乐会听众人数。

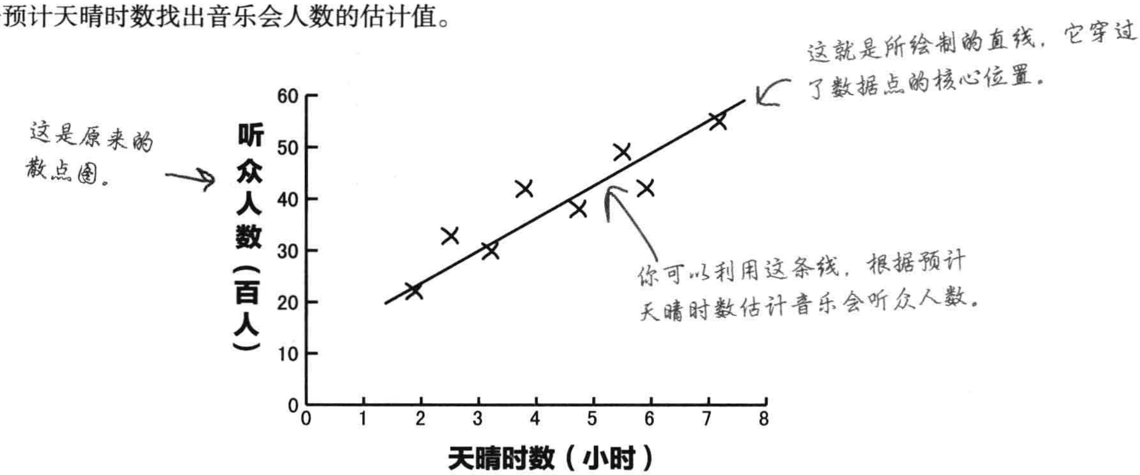
动动脑

你觉得我们该如何为二变量数据进行这类预测？

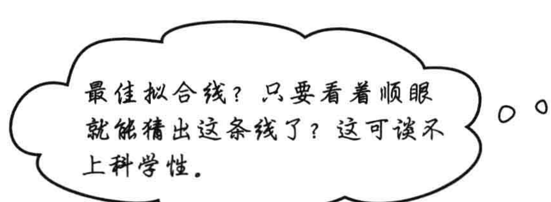
用最佳拟合线预测数值

前面讲到如何借助散点图看出是否存在某种模式，从而判定数值之间是否存在关联。那么如何利用散点图根据天晴时数预测音乐会听众人数呢？—— 在已知当天天晴时数期望值的情况下，你会如何利用现有散点图预测音乐会听众人数？

其中一个办法是，在散点图上画一条穿过这些点的直线，使这条线尽量接近各个点。你无法令这条直线穿过每一个点，不过，若存在线性相关性，则应该可以保证每一个点合理地接近你所绘制的直线。如此一来，你就可以根据预计天晴时数找出音乐会人数的估计值。



能最好地接近所有数据点的线被称为**最佳拟合线**。



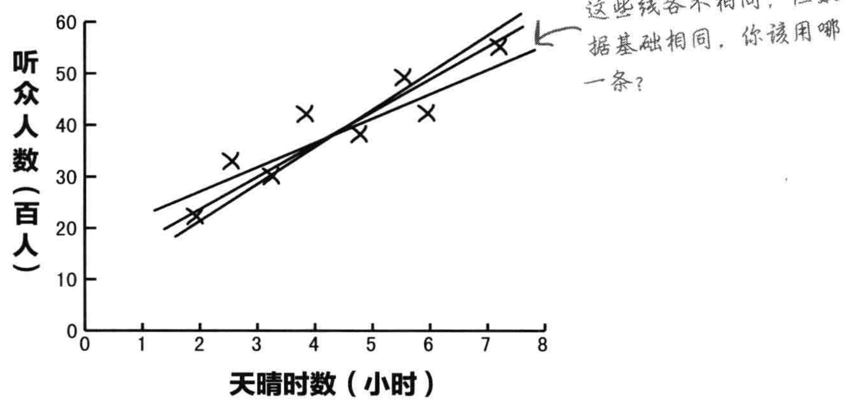
用这种方法画出的线只是一种最佳猜测。

用以上方法绘制图线的问题在于：这只是一个估计，因此根据这条线做出的任何预测都值得怀疑。你没有什么精确的方法量度这条线是否确实是最佳匹配线。这条线具有主观性，这条线的拟合质量取决于你的判断。



最佳猜测仍是猜测

假想你请三个人按照他们各自的想法画出音乐会听众人数最佳拟合线，很可能每个人都会画出与别人略有差别的最佳拟合线，如图所示：



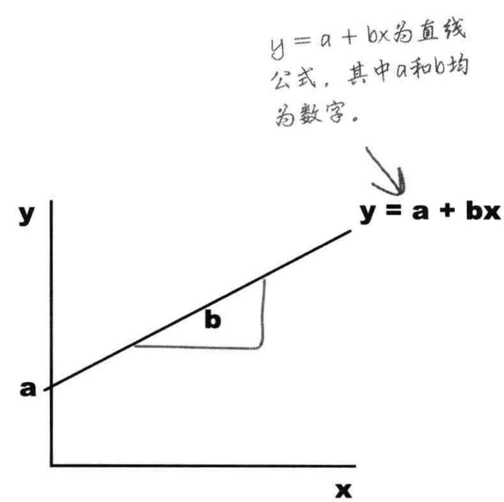
这三条线都可以想当然地被认为是数据的最佳拟合线，但我们无法知道哪一条线是名副其实的最佳拟合线。

我们确实需要找一个可以通过目测方式绘制最佳拟合线的办法。这不是一种猜测方法，而是一种更可靠的方法——使用数学或统计方法利用手头数据去找出最佳拟合线。

我们需要求出直线公式

直线的公式为y = a + bx，其中a为直线与y轴的交点，b为直线斜率，于是我们可以用公式y = a + bx表示最佳拟合线。

在我们的例子中，我们用x表示预计天晴时数，用y表示相应的露天音乐会听众人数，只要我们能用音乐会听众数据求出a和b的最合适数值，就有可靠的方法求出直线等式，且能够以更为可靠的方法按照预计天晴时数预测音乐会听众人数。

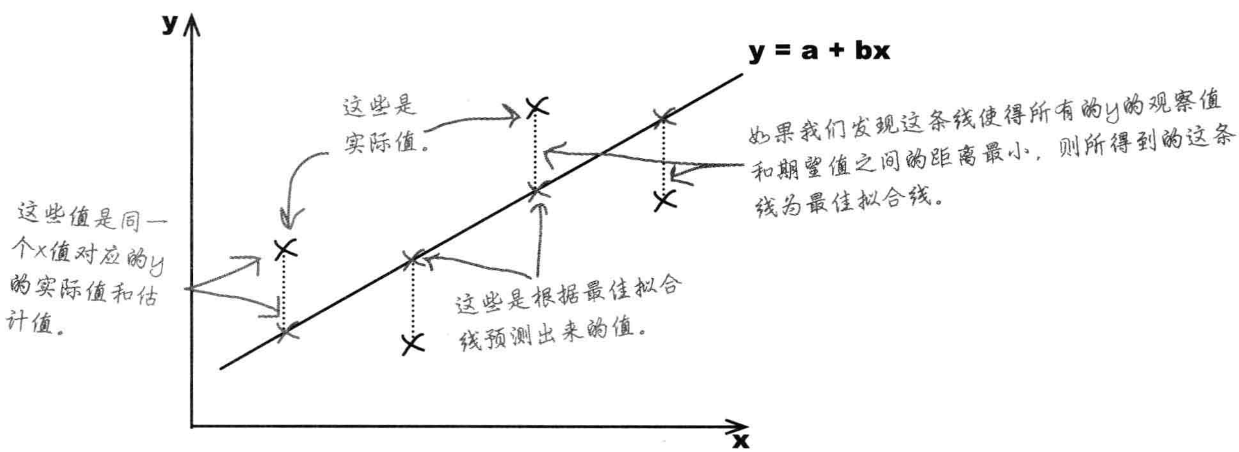


我们需要将误差最小化

让我们看看对最佳拟合线y = a + bx的要求。

最佳拟合线即能最准确地预测出所有点的真实值的线。即，对于每一个已知的x值，我们需要让数据集中的每个y变量尽可能接近我们通过最佳拟合线估计出来的数值。即，在已知某个天晴时数时，我们希望自己估计的露天音乐会听众人数尽可能接近实际值。

最佳拟合线即表达式为y = a + bx且使得y的实际观察值与每个x相对应的y的估计值的差距为最小的线。

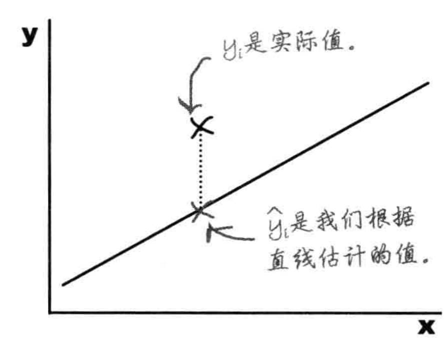


让我们用y_i表示数据集中的每一个y值，用$\hat{y}_i$表示通过最佳拟合线得出的估计值。这种表示方法与前面章节中的点估计量的表示方法一致，因为^符号代表估计值。

我们想让y的实际值和我们根据最佳拟合线得出的估计值之间的差为最小，也就是说，我们想让y_i与$\hat{y}_i$的差别之和为最小，试算如下：

$$\Sigma(y_i - \hat{y}_i)$$

可是，这个算式的问题是，实际上所有的距离都会相互抵消。我们需要稍微调整一下算法 —— 这个算法前面已经出现过了。



认识误差平方和

还记得我们第一次推导方差的时候吗？我们希望求出数据组中的数值与均值之间的距离之和，但这些距离却相互抵消。为了解决这个问题，我们将所有距离先求平方，然后加总，从而确保所有数值都是正的。

现在我们碰到了类似的情况。需要先将所有距离求平方再加总，而不是计算一对对实际值和期望值的距离之和。这样我们就能保证所有的数值都是正的。

距离平方之和被称为**误差平方和**，英文缩写为**SSE**。算式如下：

误差平方和 →

$$SS\text{-}E = \Sigma (y - \hat{y})^2$$

y的实际值以及通过最佳拟合线得出的预测值之间的差值。

即，取各个数值y，减去通过最佳拟合线得出的y的预测值，求其平方，然后将所有平方数加起来。

SSE让我们想起了方差。方差用的是数值与均值的距离的平方，SSE用的是数值与直线的距离的平方。



方差与SSE的计算方法相似。

SSE并非方差，不过，它确实涉及两个特定点之间的距离的平方——它给出了y的实际值和根据最佳拟合线得出的y的预测值之间的距离的平方之和。

我们现在需要做的就是根据y = a + bx这条线，求出使得SSE最小的a和b的数值。

求最佳拟合线公式

前面讲到我们想得到误差平方和 $\Sigma(y - \hat{y})^2$为最小的直线式，其中$y = a + bx$，从而可以得到a和b的最优值，进而得到最佳拟合线公式。

让我们先算b

$y = a + bx$中的b代表这条直线的**斜率**，或者叫陡度，即b是最佳拟合线的斜率。

我们就不进行证明了，下面直接给出使得$\Sigma(y - \hat{y})^2$ 为最小的b值：

每一个x值减去x的均值，乘以相应的y值减去y的均值

$$b = \frac{\Sigma((x - \bar{x})(y - \bar{y}))}{\Sigma(x - \bar{x})^2}$$

这有点像x的方差的算法——用每个数值x减去x的均值，然后将所得结果进行平方。



计算初看很复杂，但实际上并不那么难。

首先，求出$\bar{x}$和$\bar{y}$ —— 手头数据的x均值和y均值，此后，对每一个观察结果计算$(x - \bar{x})$ 乘以$(y - \bar{y})$，然后将结果加起来。最后，用整个结果除以$\Sigma(x - \bar{x})^2$。公式的最后一部分与样本方差的计算方法十分相似，唯一的区别是这里不除以(n-1)。你也可以利用软件完成所有计算。

下面让我们看看实际运用。



放轻松

如果在考试中需要用到这个公式，几乎可以肯定会给出这个公式。

也就是说你不用记住这个公式，只要会用就行了。

求最佳拟合线斜率

让我们看看能否用以上公式求出描述音乐会数据的直线$y = a + bx$的斜率，首先回顾一下数据：

x（天晴时数）	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y（听众人数）	22	33	30	42	38	49	42	55

让我们先求$\bar{x}$和$\bar{y}$，即x和y的样本均值。计算方法和以前完全一样，即：

$$\bar{x} = (1.9 + 2.5 + 3.2 + 3.8 + 4.7 + 5.5 + 5.9 + 7.2)/8$$
$$= 34.7/8$$
$$= 4.3375$$

利用x值求$\bar{x}$，利用y值求$\bar{y}$。

$$\bar{y} = (22 + 33 + 30 + 42 + 38 + 49 + 42 + 55)/8$$
$$= 311/8$$
$$= 38.875$$

求出$\bar{x}$和$\bar{y}$以后，就可以借助这些值用本页前一页的公式算出b。

借助$\bar{x}$和$\bar{y}$求出b

公式的第一部分是$\Sigma(x - \bar{x})(y - \bar{y})$，为此我们取各个观察结果的x值和y值，用x减$\bar{x}$，用y减$\bar{y}$，然后将两个差相乘，对每个观察结果完成以上计算以后，再将所有乘积加起来。

$x - \bar{x}$　　$y - \bar{y}$

$$\Sigma(x - \bar{x})(y - \bar{y}) = (1.9 - 4.3375)(22 - 38.75) + (2.5 - 4.3375)(33 - 38.75) + (3.2 - 4.3375)(30 - 38.75) + (3.8 - 4.3375)(42 - 38.75) + (4.7 - 4.3375)(38 - 38.75) + (5.5 - 4.3375)(49 - 38.75) + (5.9 - 4.3375)(42 - 38.75) + (7.2 - 4.3375)(55 - 38.75)$$

$(x - \bar{x})(y - \bar{y})$

$$= (-2.4375)(-16.75) + (-1.8375)(-5.875) + (-1.1375)(-8.875) + (-0.5375)(3.125) + (0.3625)(-0.875) + (1.1625)(10.125) + (1.5625)(3.125) + (2.8625)(16.125)$$

将每一组数据的所有乘积相加。

$$= 40.828125 + 10.7953125 + 10.0953125 - 1.6796875 - 0.3171875 + 11.7703125 + 4.8828125 + 46.1578125$$

$$= 122.53 \text{ (保留2位小数)}$$

求最佳拟合线的斜率，第二部分

这是公式提示。

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

下面是音乐会听众人数和预计天晴时数提示：

x（天晴时数）	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y（听众人数）	22	33	30	42	38	49	42	55

我们用各个x值求出了$\Sigma(x-\bar{x})^2$，这与样本方差的算法很相似，但不用除以(n−1)。

我们正在进行$y = a + bx$中b值的计算。我们已求得$\bar{x} = 4.3375$，$\bar{y} = 38.875$，$\Sigma(x-\bar{x})(y-\bar{y}) = 122.53$。最后要求的是$\Sigma(x-\bar{x})^2$，让我们算下去：

$$\Sigma(x-\bar{x})^2 = (1.9-4.3375)^2 + (2.5-4.3375)^2 + (3.2-4.3375)^2 + (3.8-4.3375)^2 + (4.7-4.3375)^2 + (5.5-4.3375)^2 + (5.9-4.3375)^2 + (7.2-4.3375)^2$$

$(x-\bar{x})^2$

$$= (-2.4375)^2 + (-1.8375)^2 + (-1.1375)^2 + (-0.5375)^2 + (0.3625)^2 + (1.1625)^2 + (1.5625)^2 + (2.8625)^2$$

$= 23.02$ (保留2位小数)

注意，这里不用y和$\bar{y}$。

用$\Sigma(x-\bar{x})(y-\bar{y})$ 除以 $\Sigma(x-\bar{x})^2$，即得到数值b，因此：

$$b = 122.53/23.02$$
$$= 5.32$$

我们已经求出了b，由此得到最佳拟合线的斜率。

即，数据的最佳拟合线为$y = a + 5.32x$。不过，a是多少呢？

世上没有傻问题

问： 你给出的公式看上去是针对样本的，不是针对总体的。对吗？

答： 对。我们用了样本而不用总体，这是因为我们手头有的数据是样本数据。要是你有总体数据的话，请尽管用，只要用μ代替$\bar{x}$就行了。

问： 数值b永远是正数吗？

答： 不一定。b到底是正还是负取决于线性相关类型，若为正线性相关，则b为正，若为负线性相关，则b为负。

问： 我还听说过“陡度”一词，它是什么意思？

答： 陡度是直线斜率b的另一个名称。

问： 要是不存在相关关系怎么办？我还能算出b吗？

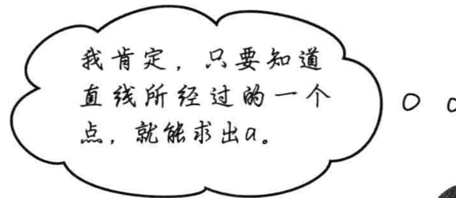
答： 如果不存在相关关系，你仍然可以通过技术手段求出最佳拟合线，但这不是数据的有效模型，无法通过这个模型做出准确预测。

问： 计算b有简便方法吗？

答： 如果观测结果很多的话，计算b十分繁琐，不过你可以借助软件进行计算。

b求出来了，a呢？

前面求出了最佳拟合线y = a + bx的最佳b值，可是我们还不知道a值。





直线需要穿过点($\bar{x}$, $\bar{y}$).

最佳拟合线最好穿过x和y的均值($\bar{x}$, $\bar{y}$)，为了确保这一点，我们用$\bar{x}$和$\bar{y}$代入直线公式y = a + bx。得到：

$$\bar{y} = a + b\bar{x}$$

或

$$\mathbf{a = \bar{y} - b\bar{x}}$$

我们已经求出了$\bar{x}$、$\bar{y}$和b的值，代入这些值，得：

$$a = 38.875 - 5.32(4.3375)$$
$$= 38.875 - 23.0755$$
$$= 15.80\ (\text{保留2位小数})$$



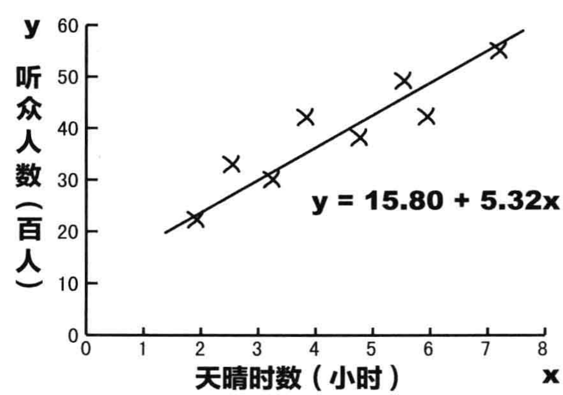
放轻松

如果你正在参加统计学考试，可能会给出这个公式。

这就是说你不必记住这个公式，只要懂得用就行了。

于是最佳拟合线公式为：

$$y = 15.80 + 5.32x$$



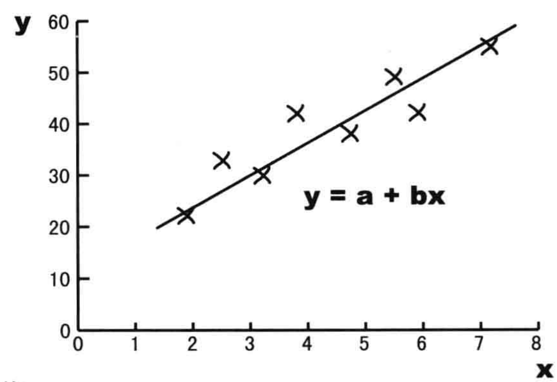
最小二乘回归法细细看

我们用于求出最佳拟合线的数学方法称为**最小二乘回归法**。

最小二乘回归法是一种数学方法，可用一条最佳拟合线将一组二变量数据拟合，通过将公式为y = a +bx的一条直线与一组数值相拟合，使得误差平方和最小——即，使得实际数值与这些数值的估计值之间的差值最小。误差平方和的公式为：

$$SSE = \Sigma(y - \hat{y})^2$$

为了对一组数据使用最小二乘回归法，需要求出a和b的值，使数据点与直线y = a + bx的拟合度最大，且SSE最小。a和b计算如下：



$$b = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

和

$$a = \bar{y} - b\bar{x}$$

求出最佳拟合线y = a + bx之后，就可以用这条线根据已知的x值预测y值，这时只要将x代入等式y = a + bx即可。

直线y = a + bx被称为**回归线**。



在预测一个特定x值对应的y值时，要避免对已知数据点范围以外的值进行预测。

线性回归法只是根据手头拥有的信息进行估计的一种方法，它体现了已知的各个数据点之间的关系，这并不表示它也适用于数据限值以外的范围。

动动笔

我们已经求出了回归线方程式，音乐会组织者在此有两个问题要请教你。下面再提示一下回归线公式：

$$y = 15.80 + 5.32x$$

其中x是预计天晴时数，y是音乐会听众人数，以“百人”为单位。

下一场音乐会当天天晴时数预计为6小时，问期望听众人数是多少？

如果音乐会听众人数会在3,500人以下，音乐会组织者将没有利润，因此将取消音乐会。问相应的预计天晴时数为多少？

动动笔解答

我们已经求出了回归线方程式，音乐会组织者在此有两个问题要请教你。下面再提示一下回归线公式：

$$y = 15.80 + 5.32x$$

其中x是预计天晴时数，y是音乐会听众人数，以“百人”为单位。

下一场音乐会当天天晴时数预计为6小时，问期望听众人数是多少？

由于x是预计天晴时数，已知x=6。我们需要求出相应的音乐会听众人数预测值，也就是要求这个x值对应的y值。

$$\begin{aligned} y &= 15.80 + 5.32x \\ &= 15.80 + 5.32 \times 6 \\ &= 15.80 + 31.92 \\ &= 47.72 \end{aligned}$$

由于y的单位为“百人”，因此期望的音乐会听众人数为47.72 × 100 = 4772。

如果音乐会听众人数会在3,500人以下，音乐会组织者将没有利润，因此将取消音乐会。问相应的预计天晴时数为多少？

这一次要求的是特定y值的相应x值。音乐会听众人数为3500，即y = 35，于是：

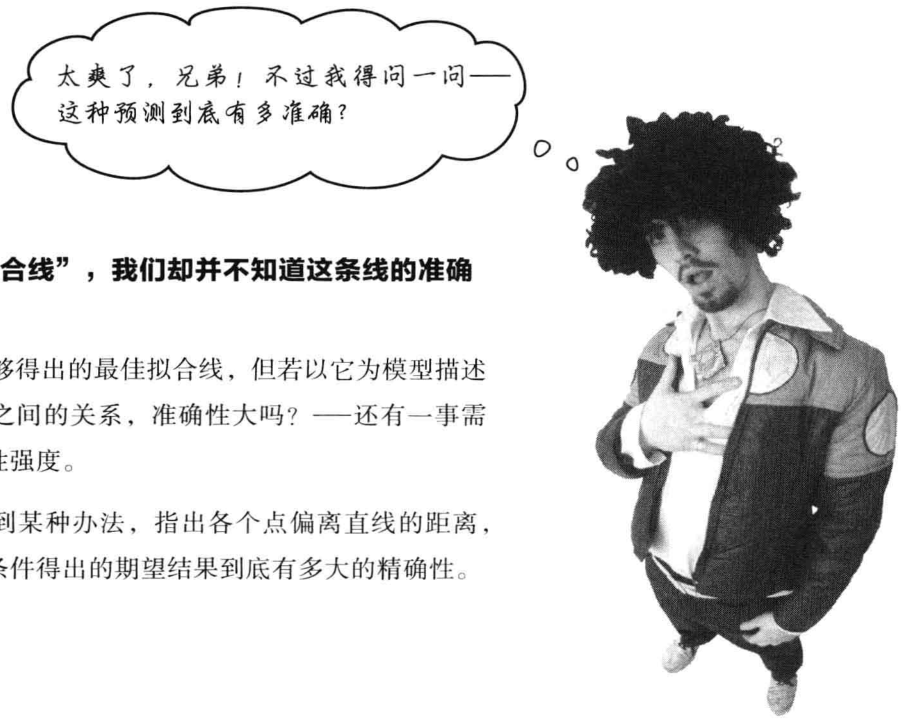
$$\begin{aligned} y &= 15.80 + 5.32x \\ 35 &= 15.80 + 5.32x \\ 35 - 15.80 &= 5.32x \\ 19.2 &= 5.32x \\ x &= 19.2/5.32 \\ &= 3.61 \text{（保留两位小数）} \end{aligned}$$

即，我们预测出的结果是：如果预计天晴时数少于3.61小时，则音乐会听众人数将低于3,500人。

你已经找出了关系

到此为止，你已经使用线性回归法建立了预计天晴时数与音乐会听众人数之间的关系模型。利用$y = a + bx$，只要知道预计天晴时数，就能预测出音乐会听众人数。

能够预测听众人数意味着你将能切实帮助音乐会组织者了解能够对票房寄予多大期望，他们还能在合理范围内期待每场演出能够实现的利润。



尽管美其名曰“最佳拟合线”，我们却并不知道这条线的准确性如何。

直线$y = a + bx$是我们能够得出的最佳拟合线，但若以它为模型描述天晴时数与音乐会听众之间的关系，准确性大吗？——还有一事需要考虑：回归线的相关性强度。

切实有用的做法是，找到某种办法，指出各个点偏离直线的距离，这会告诉我们根据已知条件得出的期望结果到底有多大的精确性。

让我们看几个例子。

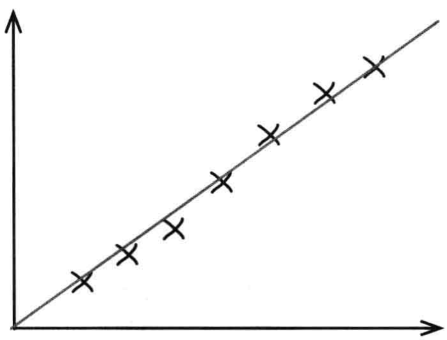
动动脑

你为什么认为了解相关强度十分重要？你觉得这会给音乐会组织者带来什么影响？

让我们查看一些相关关系

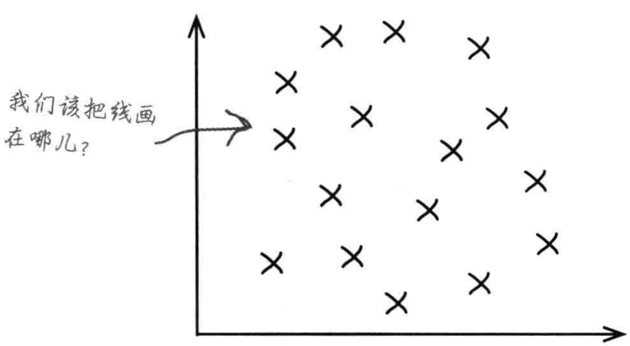
一组数据的最佳拟合线是我们所能得出的可作为两个变量之间数学关系模型的最佳直线。

尽管最佳拟合线是与数据拟合程度最高的直线，但它并不可能与每一个点都精确拟合。让我们观察几组数据，看看直线与数据的拟合情况。



精确线性相关

这一组数据的线性相关性呈现出精确的数据拟合。回归线并非百分之百完美，但几近如此。很可能依据这条线做出的任何预测都是准确的。



非线性相关

这一组数据未体现出线性相关性。你可能能用最小二乘回归法算出一条回归线，但据此做出的任何预测都不太可能准确。

你能发现问题所在吗？

两组数据都有回归线，但数据的实际拟合程度却大不相同。第一组数据的相关性十分明显，但第二组数据十分分散，以至回归线丧失应有的作用。

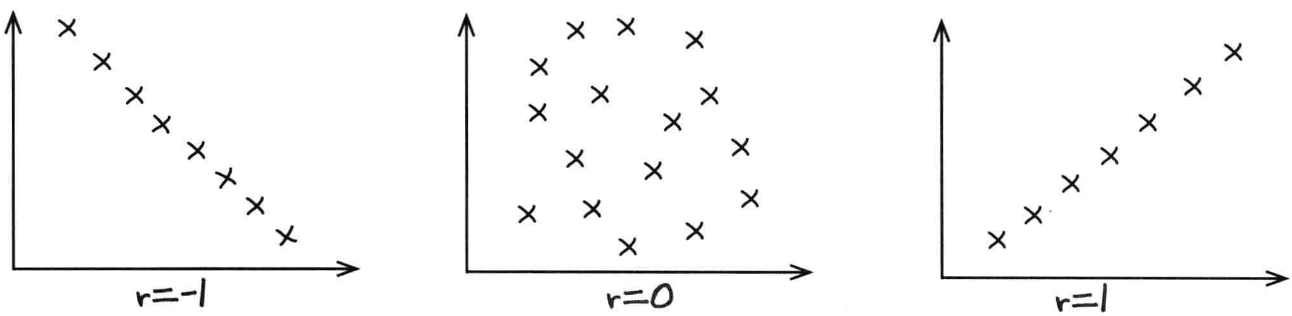
最小二乘估计可用于预测数值，也就是说，如果有某种方法能够指出数据点与直线的拟合程度，同时能指出我们的期望预测结果能够达到的精确程度，那么最小二乘估计就能发挥作用。

有一种方法可用于计算直线拟合度 —— 称为**相关系数**。

用相关系数衡量直线与数据的拟合度

相关系数是介于-1和1之间的一个数，描述了各个数据点与直线的偏离程度。通过它可以量度回归线与数据的拟合度，通常用字母r表示。

如果r等于-1，则数据为**完全负线性相关**，所有数据点都在一条直线上；如果r等于1，则数据**完全正线性相关**。如果r等于0，则**不存在相关**性。



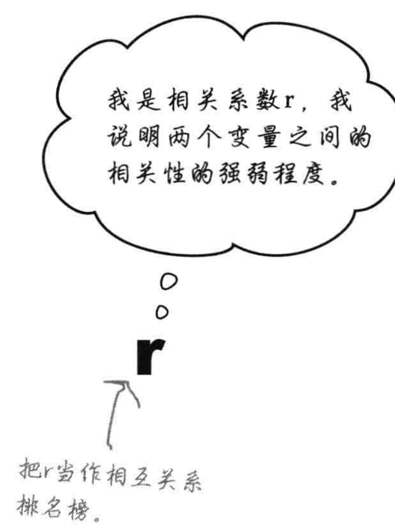
-1、0和1均为极值，通常r为介于这几个极值之间的数值。

如果r为负，则两个变量之间存在**负线性相关**。r越接近-1，相关性越强，数据点距离直线越近。

如果r为正，则两个变量之间存在**正线性相关**。r越接近1，相关性越强。

总之，随着r向0靠近，**线性相关性变弱**。于是回归线无法像r接近1或接近-1时那样准确地预测y值,数据模式可能会随机变化，或者说变量之间的关系可能是非线性的。

如果我们能算出音乐会数据的r值，就会得知我们根据预计天晴时数预测出的音乐会听众人数的准确性。如何计算r？下一页将进行讲解。



相关系数r有专用计算公式

我们如何计算相关系数r?

我们不打算在此进行证明，相关系数公式如下：

b是已求出的最佳拟合线斜率。

$$r = \frac{b\ s_x}{s_y}$$

s_x是样本中的x值的标准差，s_y是y值的标准差。

其中s_x是样本中的x值的标准差，s_y是其中y值的标准差。



我们已经完成了大部分工作。

由于我们已经算出了b，剩下的就是求S_x和S_y了。另外，我们已经完成了大部分求S_x的步骤。

在计算b的时候，我们需要求出$\Sigma(x-\overline{x})^2$的数值。如果将这个结果除以n-1，实际上就会得出x值的样本方差，如果取其平方根，则得到S_x。即：

这是样本中的x值的标准差，和以前讲过的公式相同。

$$s_x = \sqrt{\frac{\Sigma(x-\overline{x})^2}{n-1}}$$

你已经在前面算过这个值了，因此不必再算。

方程中唯一还需要计算的就是S_y——样本中的y值的标准差。其计算方法与S_x的计算方法相似：

$$s_x = \sqrt{\frac{\Sigma(y-\overline{y})^2}{n-1}}$$

这是样本中的y值的标准差，前面已经做过这一类型的计算。

让我们试着求出音乐会数据的r。

求音乐会数据的r

让我们用公式求出音乐会数据的r值。首先看一看数据提示：

x（天晴时数）	1.9	2.5	3.2	3.8	4.7	5.5	5.9	7.2
y（听众人数）	22	33	30	42	38	49	42	55

必须知道数值b、S_x及S_y才能利用本页反面的公式求出r。前面已经求出：

$$b = 5.32$$

← 这是我们先前求出的直线的斜率。

可是S_x和S_y是多少呢？

让我们先求S_x。我们先前求出$\Sigma(x-\overline{x})^2=23.02$，且已知样本大小为8。这就是说，如果我们用23.02除以7，就能得出x的样本方差。取其平方根即可得到S_x。

$$s_x = \sqrt{(23.02/7)}$$
$$= \sqrt{3.28857}$$
$$= 1.81 \text{ (保留两位小数)}$$

← 这是x值的标准差，由于是样本数值，因此除以n－1。

剩下唯一要求的就是S_y，前面已经求出$\overline{y} = 38.875$，于是：

$$\Sigma(y - \overline{y})^2 = (22 - 38.875)^2 + (33 - 38.875)^2 + (30 - 38.875)^2 + (42 - 38.875)^2 + (38 - 38.875)^2 + (49 - 38.875)^2 + (42 - 38.875)^2 + (55 - 38.875)^2$$
$$= (-16.875)^2 + (-5.875)^2 + (-8.875)^2 + (3.125)^2 + (-0.875)^2 + (10.125)^2 + (3.125)^2 + (16.125)^2$$
$$= 780.875 \text{ (保留三位小数)}$$

我们可以用以下公式求出S_y，就是将$\Sigma(y-\overline{y})^2$除以n－1，再取其平方根值。

$$s_y = \sqrt{(780.875/7)}$$
$$= \sqrt{111.55357}$$
$$= 10.56 \text{ (保留两位小数)}$$

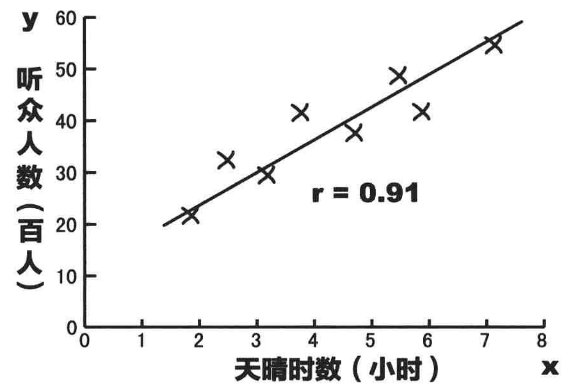
← 最后，我们用样本中的y值求出S_y——y的标准差。

现在，我们只要用b、S_x和S_y算出相关系数r就行了。

求音乐会数据的r（续）

我们已经求出b = 5.32，s_x = 1.81，s_y = 10.56，将这些结果用起来，得出r：

$$r = bs_x/s_y$$
$$= 5.32 \times 1.81/10.56$$
$$= 0.91 \text{ (保留两位小数)}$$



由于r接近1，说明露天音乐会听众人数和预计天晴时数之间有很强的正相关。换句话说，根据我们手头的数据，我们可以期望，最佳拟合线y = 15.80 + 5.32x根据预计天晴时数给出了期望音乐会听众人数的合理的良好估计。

世上没有傻问题

问： 我见过别人用其他方法计算r，他们错了吗？

答： r的计算公式有好几种形式，但这些形式本质上是一样的。我们采用的是最简单的形式，这样便于看出哪些部分已经在求b的过程中算过。

问： 这样小的一个样本能得出正确结果吗？

答： 样本大一点儿当然更好，我们用小样本只是为了让计算过程更容易看懂。

问： 你既没有证明也没有推导b和r的计算公式，为什么不做呢？

答： 推导b和r的计算公式既繁且杂，本书决定不予推导。关键是要了解使用时机、使用方法。

问： 如果预计天晴时数为0，听众人数的期望值是多少？

答： 我们无法肯定地回答这个问题，因为这已经远远超出我们的数据范围。对于在我们所拥有的数据范围以内的数据，最佳拟合线能给出相当良好的估计，但对于这个数据范围以外的数据，我们就毫无把握。那些数据可能具有其他模式，因此我们所给出的任何估计都是不可靠的。

问： 前面讲到平均数的时候，我们曾经看出单变量数据可能出现异常值。那么二变量数据呢？

答： 没错，二变量数据也可能出现异常值。异常值即距离回归线极远的那些点。如果存在异常值，则可能意味着你的数据集中有异常情况，或者，说明你的回归线与数据的拟合程度不佳。

问： 我曾经听人说起过“有影响观察结果”，这是什么东西？

答： “有影响观察结果”是一些在水平方向上与其余点相距甚远的点，因此，它们有一种将回归线朝着它们拉近的效果。.

问： 这么说有影响观察结果和异常值是一回事儿？

答： 不对。异常值远远偏离回归线，而有影响观察结果则是在水平方向上远离数据的点。

你力挽狂澜！

你对音乐会数据的计算让音乐会组织者大为惊讶，现在他们可以根据天气预报预测音乐会听众的可能人数了，也就是说有办法让利润达到最大值。



加强练习

妖怪思凡达正在采集数据——关于辐射对阿梅森上尉的超人力量产生的影响。下面是辐射时间与阿梅森上尉能够举起的吨重的成对数据。

辐射时间（分钟）	3	3.5	4	4.5	5	5.5	6	6.5	7
重量（吨）	14	14	12	10	8	9.5	8	9	6

你的任务是用最小二乘回归法求出最佳拟合线，然后求出相关系数，说明直线与数据的关联强度。请画出散点图。

如果思凡达让阿梅森上尉在辐射线下照射5分钟，你期望阿梅森上尉举起多重的重量？

纸张足够，
请尽情地算吧！

加强练习解答

妖怪思凡达正在采集数据——关于辐射对阿梅森上尉的超人力量产生的影响。下面是辐射时间与阿梅森上尉能够举起的吨重的成对数据。

辐射时间（分钟）	3	3.5	4	4.5	5	5.5	6	6.5	7
重量（吨）	14	14	12	10	8	9.5	8	9	6

你的任务是用最小二乘回归法求出最佳拟合线，然后求出相关系数，说明直线与数据的关联强度。请画出散点图。

如果思凡达让阿梅森上尉在辐射线下照射5分钟，你期望阿梅森上尉举起多重的重量？

让我们用x表示辐射时间，用y表示举起的吨重。我们需要求出回归线$y = a + bx$，因此让我们先求$\bar{x}$和$\bar{y}$.

$$\bar{x} = (4 + 4.5 + 5 + 5.5 + 6 + 6.5 + 7)/7$$
$$= 38.5/7$$
$$= 5.5$$
$$\bar{y} = (12 + 10 + 8 + 9.5 + 8 + 9 + 6)/7$$
$$= 62.5/7$$
$$= 8.9 \text{（保留两位小数）}$$

接着，让我们计算$\Sigma(x-\bar{x})(y-\bar{y})$、$\Sigma(x-\bar{x})^2$及b。

$$\Sigma(x-\bar{x})(y-\bar{y}) = (4-5.5)(12-8.9) + (4.5-5.5)(10-8.9) + (5-5.5)(8-8.9) + (5.5-5.5)(9.5-8.9) + (6-5.5)(8-8.9) + (6.5-5.5)(9-8.9) + (7-5.5)(6-8.9)$$
$$= (-1.5)(3.1) + (-1)(1.1) + (-0.5)(-0.9) + (0)(0.6) + (0.5)(-0.9) + (1)(0.1) + (1.5)(-2.9)$$
$$= -4.65 - 1.1 + 0.45 + 0 - 0.45 + 0.1 - 4.35$$
$$= -10$$

$$\Sigma(x-\bar{x})^2 = (4-5.5)^2 + (4.5-5.5)^2 + (5-5.5)^2 + (5.5-5.5)^2 + (6-5.5)^2 + (6.5-5.5)^2 + (7-5.5)^2$$
$$= (-1.5)^2 + (-1)^2 + (-0.5)^2 + 0^2 + 0.5^2 + 1^2 + 1.5^2$$
$$= 2.25 + 1 + 0.25 + 0 + 0.25 + 1 + 2.25$$
$$= 7$$

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$
$$= -10/7$$
$$= -1.43 \text{（保留两位小数）}$$

求出b后，即可用b求a。

$$a = \bar{y} - b\bar{x}$$
$$= 8.9 + 1.43 \times 5.5$$
$$= 8.9 + 7.86$$
$$= 16.76$$

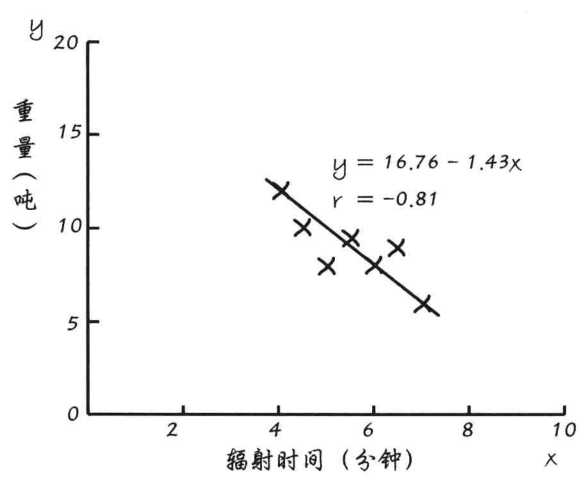
于是得出最佳拟合线为$y = 16.76 - 1.43x$。

相关系数r的计算式为$r = bs_x/s_y$，其中s_x和s_y为变量x和变量y的标准差。在求出b以后，还需要求s_x和s_y。

$$s_x = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$$
$$= \sqrt{7/6}$$
$$= 1.08$$

$$\Sigma(y-\bar{y})^2 = (12-8.9)^2 + (10-8.9)^2 + (8-8.9)^2 + (9.5-8.9)^2 + (8-8.9)^2 + (9-8.9)^2 + (6-8.9)^2$$
$$= 3.1^2 + 1.1^2 + (-0.9)^2 + 0.6^2 + (-0.9)^2 + 0.1^2 + (-2.9)^2$$
$$= 9.61 + 1.21 + 0.81 + 0.36 + 0.81 + 0.01 + 8.41$$
$$= 21.22$$

$$s_y = \sqrt{\frac{\Sigma(y-\bar{y})^2}{n-1}}$$
$$= \sqrt{21.77/6}$$
$$= 1.90$$



综合以上得：

$$r = bs_x/s_y$$
$$= -1.43 \times 1.08/1.9$$
$$= -0.81 \text{（保留两位小数）}$$

如果$x = 5$，则：

$$y = 16.76 - 1.43x$$
$$= 16.76 - 1.43 \times 5$$
$$= 9.61$$

这就是说，在辐射线下照射5分钟后，我们期望阿梅森上尉能够举起9.61吨重量。

要 点

- 单变量数据仅涉及一个变量，二变量数据涉及两个变量。
- 散点图显示出二变量数据的模式。
- 相关性是变量之间的数学关系，但并不意味着一个变量一定与另一个变量相关。线性相关即两变量间为直线的相关关系。
- 正线性相关即x的低端值对应于y的低端值，x的高端值对应于y的高端值；负线性相关即x的低端值对应于y的高端值，x的高端值对应于y的低端值。如果x和y的数值分布表现出随机模式，则它们不存在相关性。
- 与数据点拟合程度最高的线称为最佳拟合线。
- 线性回归法是一种求最佳拟合线 $y = a + bx$的数学方法。
- 误差平方和SSE的计算式为：$\Sigma(y-\hat{y})^2$。
- 直线 $y = a + bx$的斜率b的计算式为：

$$b = \frac{\Sigma(x-\bar{x})(y-\bar{y})}{\Sigma(x-\bar{x})^2}$$

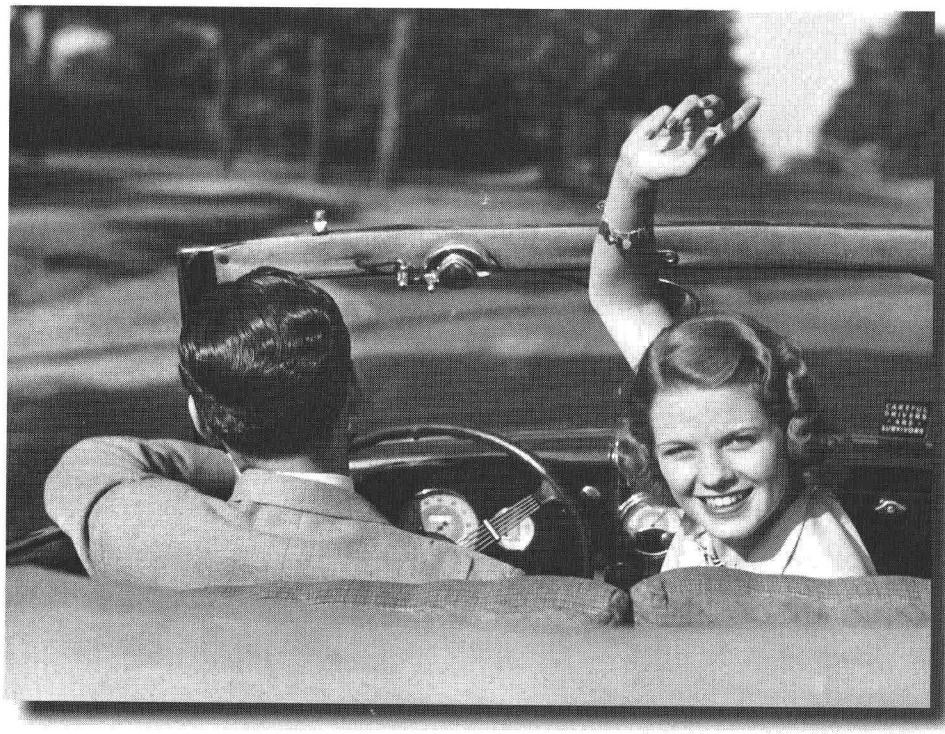
- a的计算式为：

$$a = \bar{y} - b\bar{x}$$

- 相关系数r是介于−1和1之间的一个数，描述的是数据与最佳拟合线的偏离距离。如果 $r = -1$，则为完全负线性相关；如果 $r = 1$，则为完全正线性相关；如果 $r = 0$，则不存在相关性。r的计算式为：

$$r = \frac{b\,s_x}{s_y}$$

再见……



统计邦感谢您的光临！

离别让人黯然神伤，不过，看到你能学以致用，我们真是再高兴不过了。后文尚留有不少遗珠散玉等你拾取 —— 一些方便实用的概率表、一份需要通读的索引，此后，就该把所有这些新学问付诸实践了。我们渴望知道你的消息，所以请到Head First图书馆网站（www.headfirstlabs.com）给我们**写几句**吧，让我们知道统计学为**你**做出的贡献！

附录I：尾声

正文未及的十大拓展



正文既已，余兴未尽。

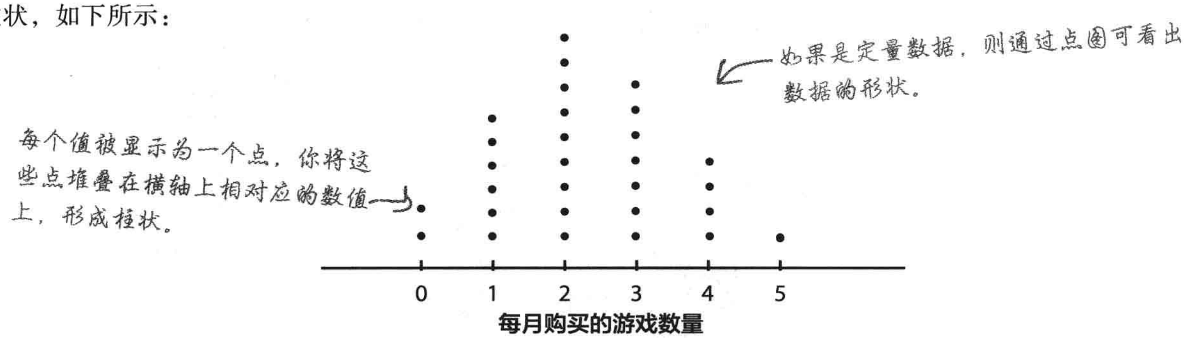
我们觉得还有一些内容是你需要知道的，对这些内容只字不提恐有不妥，不过，其实也只需要**简单地提一提** —— 我们诚挚地希望为你呈上一本厚薄适度的书，免得你为了捧起这本书学习还得先去健身中心练练臂力。因此，请先**通读一遍这里的知识点**，再合上本书。

#1. 数据的其他表现形式

我们在第一章讲过几种图形，这里再介绍两种有可能用到的图形。

点图

点图在图上以点表示各个数值，各个点在横轴上的相应数值上方堆叠成柱状，如下所示：



茎叶图

茎叶图用于体现定量数据，通常在数据集非常小的时候使用。茎叶图显示出数据集中的每一个确切值，通过它能够轻易看出数据的形状。举例如下：

16 17 22 23 23 24 25 26 26 27 28 29 29
30 31 31 32 32 33 34 34 35 36 37 37 38
39 40 41 42 42 43 43 44 45 45 49 50 50
50 51 55 58 60

这是你的原始数据。

这是根据数据画出的茎叶图。

解图密钥：10 | 6 = 16

60	0
50	0 0 0 1 5 8
40	0 1 2 2 3 3 4 5 5 9
30	0 1 1 2 2 3 4 4 5 6 7 7 8 9
20	2 3 3 4 5 6 6 7 8 9 9
10	6 7

茎叶图的外形与直方图相似，但方向颠倒了一下。

左边的数值称为**茎**，右边的数值称为**叶**，在上面的茎叶图中，茎代表十位，叶代表个位。计算原始数据中的每个数值时，用每一片叶加上这片叶的茎即可。例如这一行：

10 | 6 7

它代表两个数字：16和17。16等于叶6加上茎10；类似地，17等于叶7加上茎10。

通常会给出一个解图密钥帮你正确地理解茎叶图，此处的密钥为10 | 6 = 16。

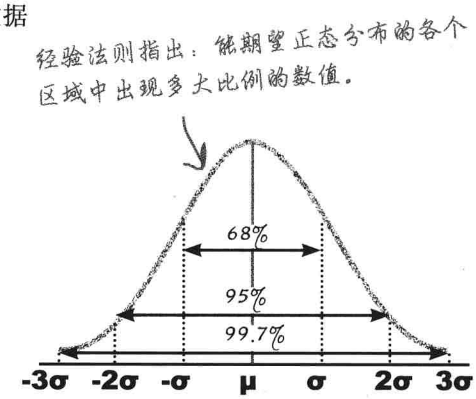
#2. 分布剖析

有两条法则可以告诉你：大部分数据落在概率分布中的哪个区域。

适用于正态分布的经验法则

经验法则适用于符合正态分布的任何数据集。它表明：几乎所有的数据都位于距离均值三个标准差的范围内。具体来说：

- **大约68%的数值位于距离均值1个标准差的范围内。**
- **大约95%的数值位于距离均值2个标准差的范围内。**
- **大约99.7%的数值位于距离均值3个标准差的范围内。**



只要知道距离均值多少个标准差就可以大致了解概率情况。

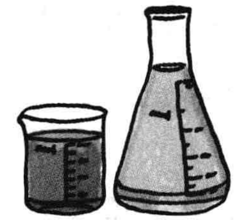
适用于任何分布的切比雪夫定理

还有一个类似的定理，它适用于任意数据集，称为**切比雪夫定理**或**切比雪夫不等式**。它指出，对于任何分布：

- **至少75%的数值位于距离均值2个标准差的范围内。**
- **至少89%的数值位于距离均值3个标准差的范围内。**
- **至少94%的数值位于距离均值4个标准差的范围内。**

切比雪夫定理不如经验法则精确，因为只给出了最小百分数，但这仍然能让你大致了解数值落在概率分布中的哪个区域。切比雪夫定理的优点是它适用于任何分布，而经验法则只适用于正态分布。

#3. 实验



实验可用于检验变量之间的因果关系。例如，通过实验可以检验不同剂量的鼾克对鼻鼾患者的治疗效果。

进行实验时，对**自变量**进行控制，以便看出对**因变量**带来的影响。例如，你可能想检验不同剂量的鼾克对患者夜间打鼾时数产生的影响。鼾克的剂量为自变量，打鼾时数则为因变量。

用于实验的对象称为**实验单位**，例中的实验单位为鼻鼾患者。

一个好实验具备哪些特点？

设计实验时要记住三个基本原则：控制（对照）、随机和重复。和抽样一样，这样做的主要目的是让偏倚达到最小值。

- **你需要对外部影响或自然变异造成的结果进行控制。**

  进行实验时，需要最小化那些不属于试验范围的影响因素。为此，我们首先要建立一个**控制组**——中文中更常叫做**对照组**，在医学试验中则为一个不进行治疗或者仅仅采用自然疗法进行治疗的中性组。通过将治疗组的治疗效果和控制组（对照组）的治疗效果进行比较，就能评估治疗效果。

  安慰剂即为一种中性疗法，它对于因变量没有影响。有时候，实验对象对中性疗法的反应与对其不进行任何治疗的反应不一样，因此，为一个组提供安慰剂是控制这种影响结果的一种办法。如果服用安慰剂的组并不知道所服用的是安慰剂，则称为**盲法**，如果连提供治疗的人也不知道这是安慰剂，则称为**双盲法**。

- **你需要将对象随机分配到采用不同疗法的治疗组中。**

  下一页详细介绍这一点。

- **你需要重复实验**

  每一种治疗方法都需要在许多对象上进行实验。鼻鼾实验需要对多位鼻鼾患者应用治疗方法，而不是只对一位患者进行治疗，这样才能评估治疗效果。

另一个要注意的问题是**混杂因素**。当一个实验的控制方法无法消除有可能对因变量造成影响的其他原因，实验就存在**混杂因素**。例如，假设你给男性服用鼾克，给女性服用安慰剂，当对这两个组的治疗结果进行比较时，就无法判断男性的治疗效果是由于药物而产生，还是由于男女两性的鼻鼾问题天生就存在差异。

实验设计

前面讲过，需要将实验对象随机分配到不同的实验组中。但如何分组最为妥当？

完全随机化设计

完全随机化设计是一种可以选用的方法。使用这种方法时，你将治疗方法完全随机地分配给实验对象。如果我们打算做一个实验检验不同剂量的鼾克对患者的治疗效果，我们会随机地把鼻鼾患者分配给特定的治疗组。例如，我们会让一半的患者服用安慰剂，另一半患者则服用鼾克。

安慰剂	鼾克
500	500

如果有1,000个对象，我们可以让一半人服用安慰剂，另一半人服用鼾克。

完全随机化设计与简单随机抽样很相似。不同的是，这里不是随机选择一个样本，而是随机分配治疗方法。

随机化区组设计

另一个可以选用的方法**随机化区组设计**。这种方法将对象划分为多个相似的**组**，或者叫做**块**，例如，你可以将鼻鼾患者分为男性组和女性组，再在每一个组内部随机分配治疗方法 —— 对于每一个性别组，可以给其中一半患者服用鼾克，给另一半患者服用安慰剂。这样做可以减小性别因素的影响，从而达到减小混杂因素的目的。

	安慰剂	鼾克
男	250	250
女	250	250

如果有500名男性和500名女性，我们会给每种性别的一半人服用安慰剂，另一半人服用鼾克。

随机化区组设计与分层随机抽样十分相似。不同的是，这里是将对象分为几个组，而不是将总体分为几个层。

配对设计

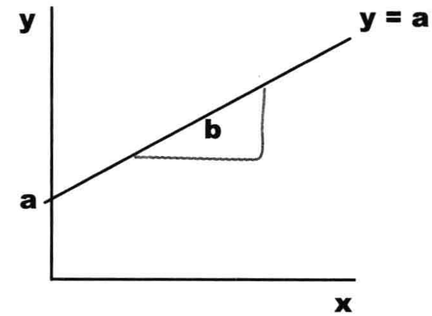
配对设计是随机化区组设计的一个特例，在只有两种治疗情况且可以将对象分为相似的对子时可以使用这种设计方法。例如，鼾克实验可以有两种治疗情况 —— 服用安慰剂或服用鼾克，而患者可以按照年龄和性别划分为相似的对子。然后，你让对子中的一位服用安慰剂，另一位服用鼻鼾。例如，如果一个对子由两名30岁的男性组成，你就可以让其中一名服用安慰剂，让另外一名服用鼾克。

	安慰剂	鼾克
男30岁	1	1
男30岁	1	1
女30岁	1	1
女30岁	1	1
...	...	...

根据年龄和性别进行配对还可以消除因为这两种因素产生的混杂因素。

#4. 最小二乘回归法的其他公式

在第15章中讲过如何求最小二乘回归线的公式y = a + bx，其中：



这是回归线的斜率公式。

$$b = \frac{\Sigma(x - \overline{x})(y - \overline{y})}{\Sigma(x - \overline{x})^2}$$

这个公式还有一种表示方法 —— 通过方差来表示，许多人觉得这种方法更便于记忆。如果：

$$s_x^2 = \frac{\Sigma(x - \overline{x})^2}{n - 1} \qquad s_y^2 = \frac{\Sigma(y - \overline{y})^2}{n - 1} \qquad s_{xy} = \frac{\Sigma(x - \overline{x})(y - \overline{y})}{n - 1}$$

数值x的样本方差　　数值y的样本方差

则回归线斜率的公式可以另行表示为：

$$b = \frac{s_{xy}}{s_x^2}$$

同一个计算式以不同方式进行表示。

类似地，可以改写相关系数的表示方法，将原来的相关系数计算式：

$$r = \frac{b\, s_x}{s_y}$$

改写为：

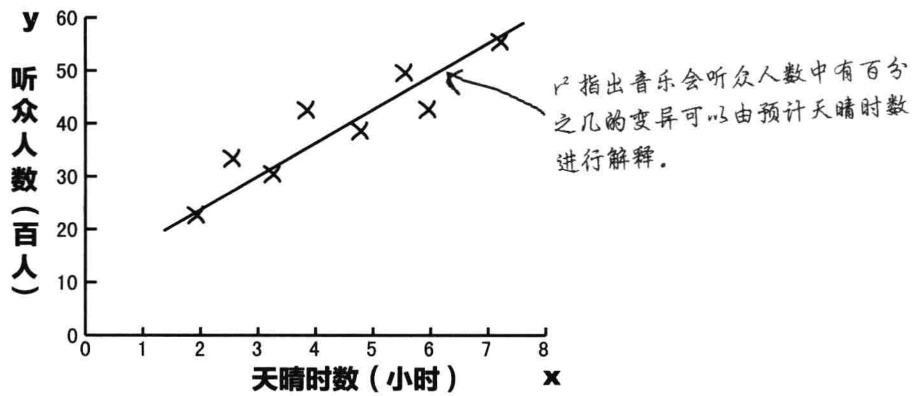
这是相关系数的公式。

$$r = \frac{s_{xy}}{s_x s_y}$$

s_{xy}称为**协方差**，正如x的方差描述x的变异情况，y的方差描述y的变异情况，x和y的协方差量度的是x和y的总变异情况。

#5. 决定系数

决定系数以r^2或R^2表示，它是可以用x变量进行解释的y变量的变异百分数。例如，可以用决定系数指出露天音乐会的听众人数中有多大比例的变异可以由预计天晴时数进行解释。



如果 $r^2 = 0$，则无法从x值预测y值。

如果 $r^2 = 1$，则可以从x值预测y值，且无误差。

通常 r^2 介于这两个极值之间，r^2 越接近1，越能通过x预测y；越接近0，r^2 越无法预测y。

计算 r^2

有两种计算r^2的方法。第一种只需要取相关系数r的平方。

这只是相关系数的平方。

$$r^2 = \left(\frac{s_{xy}}{s_x s_y}\right)^2$$

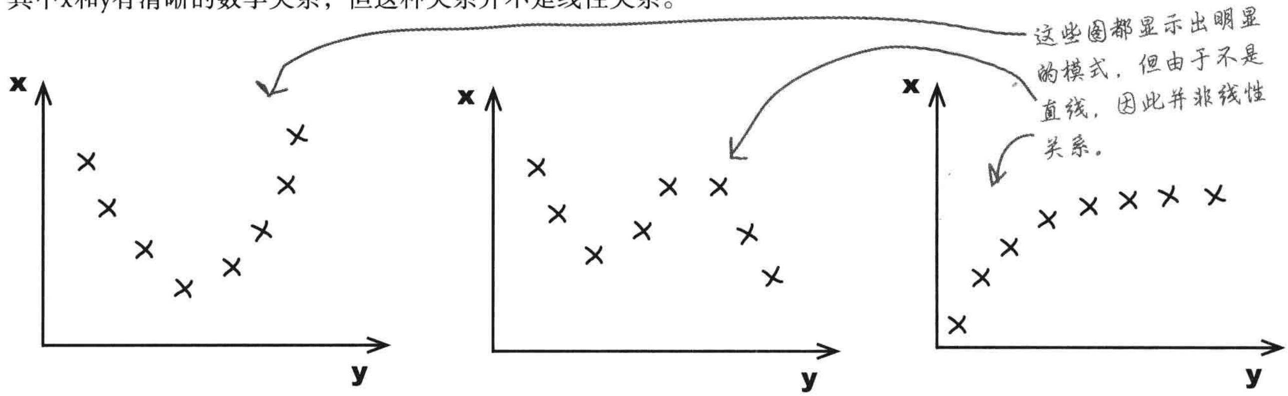
另一种方法是将各个y值与其估计值的差距取平方，然后求和，再除以y值与 $\bar{y}$ 的差距的平方的总和。

$$r^2 = \frac{\Sigma(y - \hat{y})^2}{\Sigma(y - \bar{y})^2}$$

这样计算得出的结果与上面的结果相同，只是计算方法不同而已。

#6. 非线性关系

当两个变量存在关系时，这种关系不一定是线性的。下面是一些散点图实例，其中x和y有清晰的数学关系，但这种关系并不是线性关系。



线性回归法假设两个变量之间的关系可以通过直线描述，因此对于这样的原始数据，运用最小二乘回归法无法很好地估计回归线的方程。

不过，有一个办法可以解决这个问题。有时候可以通过对x和y进行转化，使结果接近线性。然后可以对转化结果运用线性回归法，求出a和b。最大的困难在于努力将图形的非线性方程转化为以下形式：

$$y' = a + bx'$$

其中 y' 和 x' 为x的函数。

例如，你求得的最佳拟合线可能具有下列形式：

$$y = 1/(a + bx)$$

这可以变形为：

$$1/y = a + bx$$

现在符合 y' = a + bx的形式了，于是可以使用线性回归法。

于是y' = 1/y，这就是说，你可以对直线y' = a + bx运用最小二乘回归法，其中y' = 1/y。完成y值的转化后，就可以使用最小二乘回归法求出a和b的数值，然后再将结果代入原始方程。

这里讲得很简略，只是为了让你知道可能的做法。

如果最佳拟合线为非线性形式，有时候可以通过转换使其成为线性形式。

#7. 回归线斜率的置信区间

前面的章节中已经讲过如何求得 μ 和 σ^2 的置信区间，对于回归线 $y = a + bx$，还可以求出其斜率的置信区间。

b的置信区间如下：

$$\hat{b} \pm (\text{误差范围})$$

可是误差范围是多少呢？

b的误差范围

误差范围计算如下：

误差范围= t(ν) x (b的标准差)

其中 $\nu = n-2$，n为样本的观察结果数目。为了求出t(ν)的数值，可用t分布概率表查找 ν 和置信水平。

b的抽样分布的标准差计算如下：

这是b的抽样分布的标准差。

$$s_b = \frac{\sqrt{\dfrac{\Sigma(y-\hat{y})^2}{n-2}}}{\sqrt{\Sigma(x-\overline{x})^2}}$$



放轻松

参加统计学考试时，如果要用到S_b，这个公式是会被给出的。

也就是说你不用记住这个公式，只要会用就行了。

计算过程是，用y的观察值减去从回归线得出的y的估计值，所得的差进行平方，将所有的平方数加起来；然后除以n-2；取平方根；再用所得结果除以x的观察值与 $\overline{x}$ 之差的平方之和。

于是得出置信区间为：

你使用的是t分布，自由度为n-2。

$$(\hat{b} - t(v)\,s_b,\ \hat{b} + t(v)\,s_b) \qquad v = n - 2$$

求出b的标准差还有别的用处。例如，还可以用于假设检验，检验一条回归线的斜率是否具有特定值。

#8. 抽样分布 — 两个均值之间的差异

有时候，知道抽样分布的情况对于了解两个正态分布总体的均值之差十分有用，你可能想用这个差值构建一个置信区间或进行一个假设检验。例如，你可能想基于“两个正态分布的总体的均值相等”这一假设进行一个假设检验。

如果 $X \sim N(\mu_x, \sigma_x^2)$，$Y \sim N(\mu_y, \sigma_y^2)$，其中X和Y相互独立，则 $\overline{X} - \overline{Y}$ 的分布的期望和方差的计算式为：

$$E(\overline{X} - \overline{Y}) = \mu_x - \mu_y$$

这是因为 $E(\overline{X} - \overline{Y}) = E(\overline{X}) - E(\overline{Y})$

$$Var(\overline{X} - \overline{Y}) = \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}$$

类似地，$Var(\overline{X} - \overline{Y}) = Var(\overline{X}) + Var(\overline{Y})$

如果已知总体方差 σ_x^2 和 σ_y^2，则 $\overline{X} - \overline{Y}$ 符合正态分布，即：

$$\overline{X} - \overline{Y} \sim N\left(\mu_x - \mu_y, \ \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

通过上式可以求出 $\overline{X} - \overline{Y}$ 的置信区间。置信区间公式为(统计量) ± (误差范围)，因此在本例中，置信区间为：

$$\bar{x} - \bar{y} \pm c\sqrt{Var(\overline{X} - \overline{Y})}$$

这是 $\overline{X} - \overline{Y}$ 的置信区间。

c值取决于置信区间所要求的置信水平：

置信水平	c值
90%	1.64
95%	1.96
99%	2.58

c值由置信水平决定

如果 σ_x^2 和 σ_y^2 未知，则需要用 s_x^2 和 s_y^2 进行近似。如果样本很大，则仍然可以使用正态分布。如果样本很小，则需要使用t分布。

#9. 抽样分布 — 两个比例之间的差异

还有一个针对两个二项分布总体的比例差异的抽样分布，利用这个分布可以构建一个置信区间或进行一个假设检验。例如，你可能想基于“两个总体比例相等”这一假设进行一个假设检验。

如果 $X \sim B(n_x, p_x)$，$Y \sim B(n_y, p_y)$，其中X和Y相互独立，则分布 $P_x - P_y$ 的期望和方差为：

$$E(P_x - P_y) = p_x - p_y$$

像前面一样，$E(P_x - P_y) = E(P_x) - E(P_y)$

$$Var(P_x - P_y) = \frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}$$

$Var(P_x - P_y) = Var(P_x) + Var(P_y)$

如果每个总体的np和nq都大于5，则 $P_x - P_y$ 可以近似于正态分布。即：

$$P_x - P_y \sim N\left(p_x - p_y, \frac{p_x q_x}{n_x} + \frac{p_y q_y}{n_y}\right)$$

通过这个分布可以求出 $P_x - P_y$ 的置信区间。置信区间等于(统计量) ± (误差范围)，因此，在本例中，置信区间为：

$$p_x - p_y \pm c\sqrt{Var(P_x - P_y)}$$

这是 $P_x - P_y$ 的置信区间

c值取决于置信区间所要求的置信水平，c值与下一页的结果相同。



放轻松

参加统计学考试时，如果要用到两个均值或两个比例的抽样分布，是会给出抽样分布的方差的。

也就是说你不用记住这个公式，只要会用就行了。

#10. 连续概率分布的E(X)和Var(X)

在求**离散**概率分布的期望和方差时，我们使用下列算式：

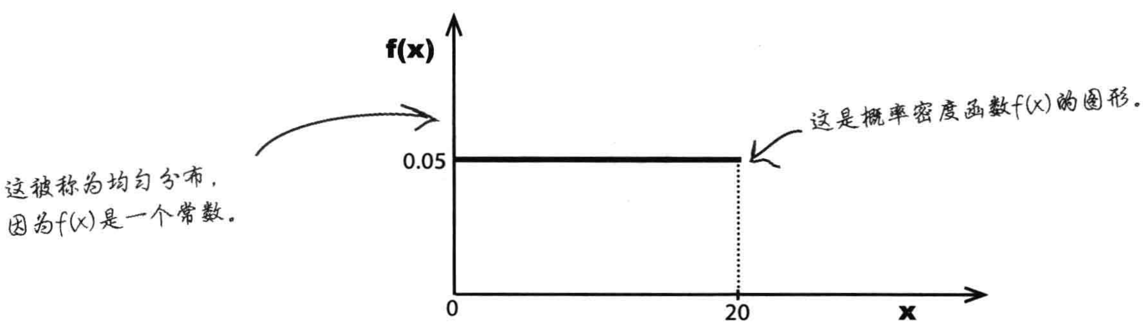
$$E(X) = \Sigma xP(X = x)$$

$$Var(X) = \Sigma x^2P(X = x) - E^2(X)$$

在概率分布为**连续分布**的时候，则通过面积求期望和方差。

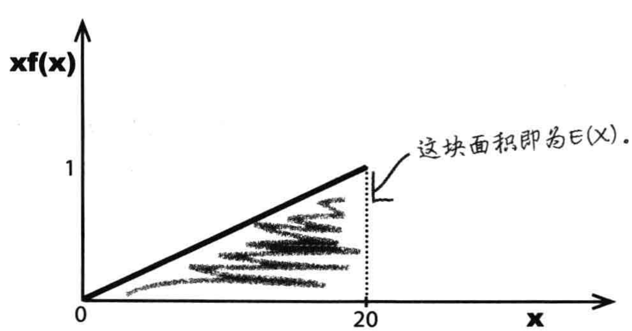
例如，假如你有一个连续概率分布，其概率密度函数如下：

$$f(x) = 0.05 \qquad 0 \le x \le 20$$



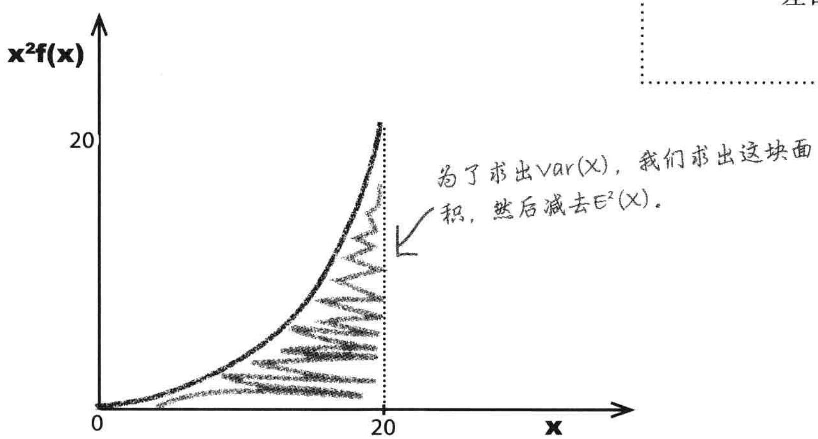
求E(X)

为了求出期望，我们需要求出概率分布范围内的曲线xf(x)下方的面积。实例中要求的是x在0至20范围内的图形0.05x下方的面积。



求Var(X)

为了求出方差，需要求出曲线$x^2f(x)$下方的面积，然后减去$E^2(X)$，即，要求出曲线$0.05x^2$下方介于0和20之间的面积，然后减去E(X)的平方。



放轻松

你并不经常需要求一个连续随机变量的期望和方差。

你经常用到的是正态分布之类，若是这种情况，期望和方差都会被给出。

通常，在整个x范围内，连续随机变量的期望和方差计算如下：

$$E(X) = \int xf(x)dx$$

$$Var(X) = \int x^2f(x)dx - E^2(X)$$

求连续随机变量的期望和方差常需用到微积分。

[市场营销部捎话：能给《深入浅出微积分》做个广告吗－很快就出版。]

重要统计量

均匀分布

如果X符合均匀分布，则：

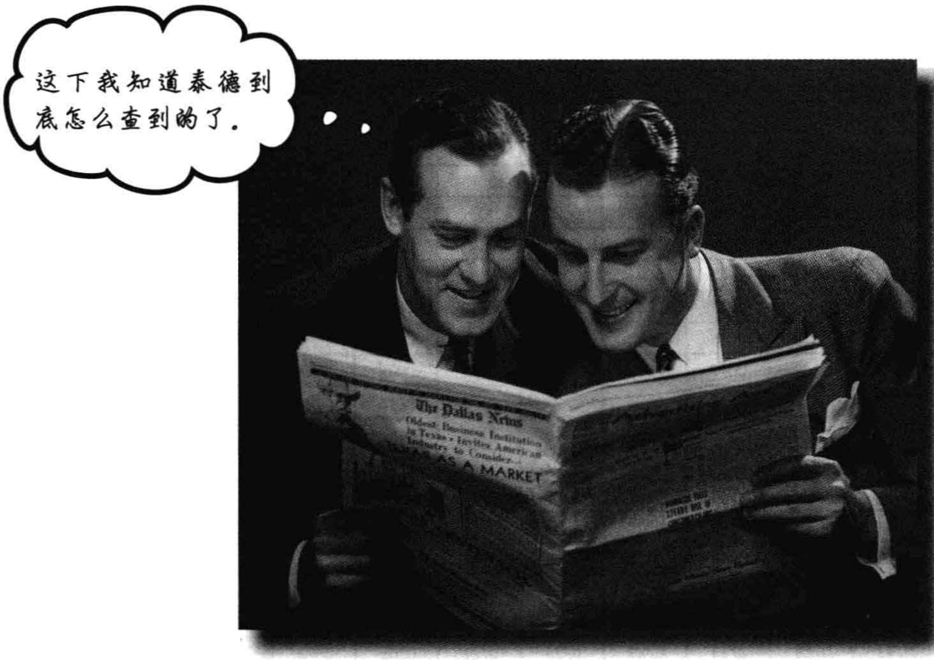
$f(x) = 1/(b - a)$ 其中 $a \le x \le b$

$E(X) = (a + b)/2$

$Var(X) = (b - a)^2/12$

附录II：统计表

快来查表

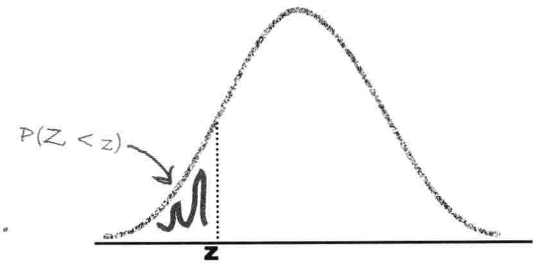


缺少值得信赖的概率表该怎么办？

仅仅了解概率分布是不够的，有时还需要在标准**概率表**中**查找概率**。这份附录给出了**正态分布**、**t分布**和χ^2**分布**的概率表，可在其中尽情查找各种概率。

#1. 标准正态分布表

通过本表可求P(Z < z)的概率，其中Z ~ N(0, 1)。为了求出P(Z < z)，可查精确到2位小数的z值，然后读出概率即可。



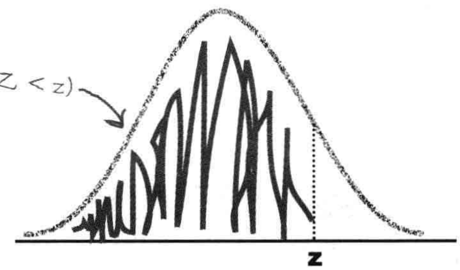
根据第一列和第一行查出z值……

……然后从表中读出概率。

这些是z为负数时P(Z < z)的概率。

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
-3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
-3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
-3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
-3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
-2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
-2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
-2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
-2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
-2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
-2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
-2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
-2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
-2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
-2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
-1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
-1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
-1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
-1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
-1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
-1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
-1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
-1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
-1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
-1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
-0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
-0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
-0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
-0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
-0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
-0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
-0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
-0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
-0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
-0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

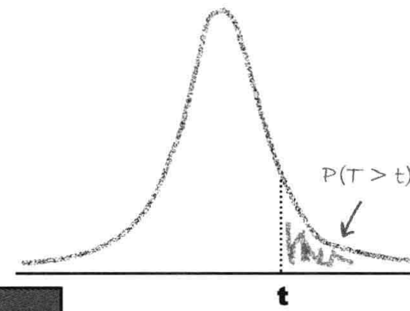
#1. 标准正态分布表（续）



这些是z为正数时P(Z < z)的概率。

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

#2. t分布临界值



本表可查出P(T > t) = p时的t值。T符合t分布，v为自由度。查找v值和p值，然后读出t。

在第一列中查找 v ……

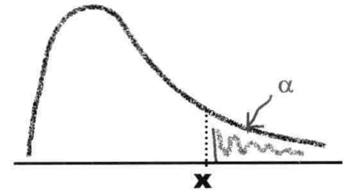
……在第一行中查找p……

然后从表中读出t的值。

t	尾部概率p											
v	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	置信水平c											

#3. X^2 临界值

本表可查出 P(X ≥ x) = α时的X值。X符合 χ^2 分布，自由度为ν。查找ν和α的值，即可读出x。



在第一列查找ν的值……

……在第一行查找α的值……

……然后从表中读出x。

	尾部概率 α										
ν	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.51
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61
80	88.13	90.41	93.11	96.58	101.9	106.6	108.1	112.3	116.3	120.1	124.8
100	109.1	111.7	114.7	118.5	124.3	129.6	131.1	135.8	140.2	144.3	149.4

索引

符号

| 符号（参见条件概率）
∩ 交集
  求解 159
  P(A ∩ B) 与 P(A | B) 165
  P(黑 ∩ 偶) 167
    P(偶) 167
1/p，期望 281
λ
  大的时候 407
  小的时候 407
λ 分布（参见泊松分布）
μ（缪） 50, 445
  confidence intervals（置信区间） 498
ν（纽） 573
  degrees of freedom（自由度） 574
Σ（西格玛） 49
  mean（均值） 49
σ（西格玛） 107
χ^2（卡方） 576
χ^2（卡方）分布 567–604
  cheat sheet（小抄） 584
  contingency table（列联表） 587
  defined（定义） 572
  degrees of freedom（自由度） 574, 576, 595
    calculating 591（计算）
    generalizing（归纳） 596–597
  expected frequencies（期望频数） 587–588
  goodness of fit（拟合优度） 573, 579, 584
  independence（独立性） 573, 586
  main uses（主要用途） 573
  significance（显著性） 575
  ν（纽） 573
χ^2（卡方）假设检验步骤 576
χ^2（卡方）概率表 575
χ^2（卡方）检验 571
$\bar{x}$（x 拔） 445–447, 472–476
  distribution of（分布） 476–486

A

accurate linear correlation（精确线性相关） 630
alternate hypothesis（备择假设） 529–530, 543
average（平均值） 46–82
  mean（均值，参见“mean”）
  median（中位数，参见“median”）
  mode（众数，参见“mode”）
  types of（类型……） 71
average distance（平均距离） 105
  interquartile range（四分位距） 105

B

bar charts（条形图） 10–20, 23
  frequency scales（频数刻度） 13
  percentage scales（百分数刻度） 12
  scales（刻度） 23
  segmented bar chart（分段条形图） 14
  split-category bar chart（分立条形图） 14
Bayes' Theorem（贝叶斯定理） 173, 178–179
bias（偏倚） 423–426, 434, 438
  in sampling（抽样……） 424–426, 438
    sources（来源） 425
bimodal（双峰） 73
binomial distribution（二项分布） 289, 324, 384, 392–393, 544
  approximating（近似） 389, 398, 407
  approximating with normal distribution（近似正态分布） 386
  approximating with Poisson distribution（近似泊松分布） 316–317
  central limit theorem（中心极限定理） 482
binomial distribution (continued)（二项分布（续））
  discrete（离散） 395

expectation and variance（期望与方差） 298, 301
finding mean and variance（求均值与方差） 389
guide（指南） 302
versus normal distribution（……与正态分布） 393, 395
Binomial Distribution Up Close（二项分布细细看） 297
binomial probabilities（二项分布概率） 384
bivariate data（二变量数据） 608, 616, 640
visualizing（图形化） 609
blinding（盲法） 646
box and whisker diagrams（箱线图） 100–102
box plot（箱形图） 100
Bullet Points（要点）
bias（偏倚） 438
binomial distribution（二项分布） 324
bivariate data（二变量数据） 640
box and whisker diagram（箱线图） 102
cluster sampling（整群抽样） 438
continuity correction（连续性修正） 396
continuous data（连续数据） 337
continuous probability distributions（连续概率分布） 337
correlation coefficient（相关系数） 640
critical region（拒绝域） 539
cumulative frequency（累积频数） 42
discrete data（离散数据） 337
expectation and variance of X（X的期望和方差） 485
expectation of random variable X（随机变量X的期望） 224
expectations（期望） 220, 233
frequency density（频数密度） 30
geometric distribution（几何分布） 324
histograms（直方图） 30
hypothesis tests（假设检验） 539
Type I error（第一类错误） 566
Type II error（第一类错误） 566
independent observations（独立观察结果） 378
independent observations of X（X的独立观察结果） 233
independent random variables（独立随机变量） 233
interpercentile range（百分位距） 102
interquartile range（四分位距） 97
kth percentile（第k百分位数） 102
linear regression（线性回归） 640
linear transforms（线性变换） 220, 224, 233
line of best fit（最佳拟合线） 640
negative linear correlation（负线性相关） 640
normal distribution（正态分布） 359
approximating（近似） 396
normal probabilities（正态概率） 359
one-tailed tests（单尾检验） 539
p-value（p值） 539
percentiles（百分位数） 102
point estimator（点估计量） 447
Poisson distribution（泊松分布） 324, 412
population（总体） 438
positive linear correlation（正线性相关） 640
probability distributions（概率分布） 220, 224
quartiles（四分位数） 97
range（距） 97
sample（样本） 438
sampling distribution of means（均值的抽样分布） 485
sampling distribution of proportions（比例的抽样分布） 466
scatter diagrams（散点图） 640
significance level（显著性水平） 539
simple random sampling（简单随机抽样） 438
standard deviation（标准差） 122, 220
σ 224
standard error of proportion（比例标准误差） 466
standard error of the mean（均值标准误差） 485
standard scores（标准分） 122
stratified sampling（分层抽样） 438
sum of squared errors（误差平方和） 640
systematic sampling（系统抽样） 438
test statistic（检验统计量） 539
two-tailed tests（双尾检验） 539
univariate data（单变量数据） 640
upper and lower bounds（上界和下界） 97
variance of random variable X（随机变量X的方差） 224
variances（方差） 122, 220, 233
z-scores（z分） 122
χ^2 distribution（χ^2分布） 598
goodness of fit test（拟合优度） 598
test for independence（检验独立性） 598

C

categorical data（类别数据） 18, 73
mean（均值） 62
median（中位数） 62
categories versus numbers（类别与数字） 18–23
causation versus correlation（因果与相关） 614
census（普查） 418
central limit theorem（中心极限定理） 481–482, 485
binomial distribution（二项分布） 482

Poisson distribution（泊松分布） 482
central tendency（集中趋势） 45–82
charts and graphs（图表） 4
bar charts（条形图） 10–20, 23
bar chart scales（条形图刻度） 23
choosing right one（做出正确选择） 39–40
comparing（比较） 6
cumulative frequency（累积频数） 35, 42
failure（遇挫） 9
frequency（频数） 8–9, 23
frequency scales（频数刻度） 13
histograms（直方图，参见"histograms"）
horizontal bar charts（水平条形图） 11, 23
line charts（线形图） 41, 42
multiple sets of data（多批数据） 14, 23
numerical data（数字数据） 23
percentage sales（百分数刻度） 12
pie charts（饼图） 8–9, 9, 23
proportions（比例） 9
scales（刻度） 12
segmented bar chart（分段条形图） 14
software（软件） 6
split–category bar chart（分立条形图） 14
vertical bar charts（垂直条形图） 10–11, 23
Chebyshev's inequality（切比雪夫不等式） 645
chi square (χ^2)（卡方(χ^2)） 576
chi square (χ^2) distribution（卡方(χ^2)分布） 567–604
cheat sheet（小抄） 584
contingency table（列联表） 587
defined（定义） 572
degrees of freedom（自由度） 574, 576, 595
calculating（计算） 591
generalizing（归纳） 596–597
expected frequencies（期望频数） 587–588
goodness of fit（拟合优度） 573, 579, 584
independence（独立性） 573, 586
main uses（主要用途） 573
significance（显著性） 575
ν（纽） 573
chi square (χ^2) hypothesis testing steps（卡方(χ^2)假设检验步骤） 576
chi square (χ^2) probability tables（卡方(χ^2)概率表） 575
chi square (χ^2) test（卡方(χ^2)检验） 571
clustered sampling（整群抽样） 434
cluster sampling（整群抽样） 433–434, 436, 438
coefficient of determination（决定系数） 649
combinations（组合，参见"permutations and combinations"）
combined weigh（综合体重）
continuous（连续） 365
distributed（分布） 367
distributed normally（正态分布） 365
complementary event（对立事件） 136
completely randomized design (experiments)（完全随机化设计） 647
conditional probabilities（条件概率） 157–160
Bayes' Theorem（贝叶斯定理） 173
P(A ∩ B) 与 P(A | B) 165
P(黑 | 偶) 170
probability tree（概率树） 158–161
confidence intervals（置信区间） 487–520, 539
cheat sheet（小抄） 504
confidence level changes（置信水平改变） 518
four steps for finding（求解置信区间四步骤） 491–502
Step 1: Choose your population statistic（第1步：选择总体统计量） 492, 508
Step 2: Find its sampling distribution（第2步：求出其抽样分布） 492, 509
Step 3: Decide on the level of confidence（第3步：决定置信水平） 494, 512
Step 4: Find the confidence limits（第4步：求出置信上下限） 496–501, 513
introducing（认识置信区间） 490
point estimators（点估计量） 493
selecting appropriate confidence level（选择合适的置信区间） 495
size of sample changes（改变样本大小） 518
confidence intervals (continued)（置信区间（续））
slope of regression line（回归线斜率） 651
summary（总结） 503
t–distributions（t分布） 509–515
probability tables（概率表） 513
shortcuts（简明表示） 515
small sample（小样本） 510
standard score（标准分） 511
versus confidence level（关于置信水平） 507
confidence level versus confidence interval（置信水平与置信区间） 507
confidence limits（置信上下限） 496, 502, 513

confounding（混杂因素） 646
contingency table（列联表） 587
continuity correction（连续性修正） 395–398, 412
Continuity Corrections Up Close（连续性修正细细看） 397
continuous data（连续数据） 327, 337, 365
frequency（频数） 328
probability distribution（概率分布） 329–333
range of values（数值范围） 333
versus discrete data（关于离散数据） 366
continuous probabilities（连续概率） 333
continuous probability distributions（连续概率分布） 337
E(X)和Var(X) 654–655
continuous random variables（连续随机变量） 331
continuous scale versus discrete probability distribution（连续刻度与离散概率分布） 395
control group（控制组（对照组）） 646
controls（控制（对照）） 646
correlation and regression（相关与回归） 605–642
accurate linear correlation（精确线性相关） 630
bivariate data（二变量数据） 608, 616, 640
visualizing（图形化） 609
correlation coefficient（相关系数） 630–634, 640
correlation versus causation（相关与因果） 614
dependent variable（因变量） 608
explanatory variable（解释变量） 608
independent variable（自变量） 608
least squares regression（最小二乘回归） 626
linear regression（线性回归） 626, 640
line of best fit（最佳拟合线） 618, 624, 640
finding equation（求公式） 622
finding slope（求斜率） 623–624
sum of squared errors（误差平方和） 620–621
negative linear correlation（负线性相关） 613, 631, 640
no correlation（不相关） 613, 631
no linear correlation（非线性相关） 630
outliers（异常值） 634
perfect negative linear correlation（完全负线性相关） 631
perfect positive linear correlation（完全正线性相关） 631
positive linear correlation（负线性相关） 613, 631, 640
regression line（回归线） 626
response variable（反应变量） 608
scatter diagrams（散点图） 609, 612, 616, 618, 640
scatter plots（散点图） 609
sum of squared errors（误差平方和） 640
univariate data（单变量数据） 608, 640
correlation coefficient（相关系数） 631–634, 640
formula（公式） 632
least square regression（最小二乘回归） 648
critical region（拒绝域） 531–534, 539, 548
Critical Regions Up Close（拒绝域细细看） 534
critical value（临界值） 532
cumulative frequency（累积频数） 34–38, 42
graph（图） 35

D

data（数据）
categorical and numerical data（类别数据与数字数据） 18
categorical data（类别数据） 18
grouped（分组） 19
multiple sets of data（多批数据） 14
numerical data（数字数据） 18
qualitative data（定性数据） 18
deciles（十分位数） 98
degrees of freedom（自由度） 574, 576, 595
calculating（计算） 591
generalizing（归纳） 596–597
number of（数量） 510
dependent events（独立事件） 181, 189–190
dependent variables (experiments)（因变量） 608, 646
discrete data（离散数据） 329, 337, 370
versus continuous data（……与连续数据） 326–327, 366
discrete probability distributions（离散概率分布） 197–240
expectation（期望） 204–208
linear transforms（线性变换） 233
expectations（期望） 219
independent observations（独立观察结果） 224, 225–226
linear relationship between E(X) and E(Y)（E(X)和E(Y)之间的线性关系） 217–218
linear transforms（线性变换） 219, 225–226
expectation and variance（期望和方差） 233
linear transforms versus playing multiple games（线性变换与多局赌博） 221
observation（观察值） 222–224
observation shortcuts（观察值速算法） 223

Pool Puzzle（奇妙池） 215–216
random variables（随机变量）
adding（增加） 230
independent（独立） 233
subtracting（减小） 231
shortcut or formula（简便算法或公式） 236
variance（方差） 205–208, 219
linear transforms（线性变换） 233
versus continuous scale（……与连续刻度） 395
discrete random variables（离散随机变量） 202
distribution（分布）
anatomy（剖析） 645
mean（均值） 56
of X + Y（X + Y…） 370
dotplots（点图） 644
double blinding（双盲法） 646
drawing lots（抽签） 431, 434

E

E(X) and Var(X) for continuous probability distributions（连续概率分布的E(X)和Var(X)） 654–655
empirical rule for normal distribution（正态分布经验法则） 645
estimating populations and samples（总体和样本的估计） 441–486
central limit theorem（中心极限定理） 481–482, 485
binomial distribution（二项分布） 482
Poisson distribution（泊松分布） 482
distribution of P_s（P_s分布） 464–466
expectation of P_s（P_s的期望） 462
formulas（公式） 451
point estimators（点估计量） 443–447, 452
for population variance（总体方差的……） 457
sampling distributions（抽样分布） 485
population mean（总体均值） 443, 446
population parameters（总体参数） 444
population proportion（总体比例） 454–457
population variance（总体方差） 448–450
probabilities for a sample（样本概率） 459
proportions, sampling distribution of（比例，抽样分布） 460
sample mean（样本均值） 445, 446
sample variance（样本方差） 449, 452
sampling distribution（抽样分布） 466
continuity correction（连续性修正） 469
of proportions（比例……） 460
sampling distribution of means（均值的抽样分布） 471–479
distribution of x（X的分布） 480
expectation for X（X的期望） 474–475
variance of X（X的方差） 476
standard error（标准误差） 485
of mean（均值……） 479
of proportion（比例……） 466
variance of Ps（Ps的方差） 463
x bar（x拔） 445
μ 445
events（事件） 132
complementary（对立） 136
dependent（独立） 181
exclusive（互斥） 147–154
versus exhaustive（穷举） 150
independent（独立） 182–184
versus dependent（…与独立） 189–190
intersecting（相交） 147–154
mutually exclusive（互斥） 147, 150
exclusive events（互斥事件） 147–154, 150
exhaustive（穷举） 149
exhaustive events（穷举事件） 150
expectations（期望） 204–208, 219, 220, 367
1/p 281
binomial distribution（二项分布） 298
expectations (continued)（期望（续））
geometric distribution（几何分布） 280–281
independent observations（独立观察结果） 378
linear transforms（线性变换） 233
Poisson distribution（泊松分布） 308
two games（两局赌局） 222–224
experimental units（实验单位） 646
experiments（实验） 646
designing（设计） 647
explanatory variable（解释变量） 608

F

factorials（阶乘） 246, 248
Fireside Chats, Dependent and Independent discuss their differences（面对面：相关与独立差异谈） 186–187
Five Minute Mystery（五分钟推理）

Case of the Broken Cookies（破碎的饼干） 315
Solved（破解） 318
Case of the High Sunscreen Sales（防晒霜销量案） 611
Solved（破解） 615
Case of the Lost Coffee Sales（消失的咖啡销量） 421
Solved（破解） 429
Case of the Missing Parameters（缺失参数案件） 357
Solved（破解） 358
Case of the Moving Expectation（活动期望案例） 211
Solved（破解） 220
The Case of the Ambiguous Average（案例：含含糊糊的平均值） 51
Solved（破解） 81
The Case of the Two Classes（瑜伽班与游泳班案例） 185
Solved（破解） 188
formulas for arrangements（排位方式的计算公式） 248
frequencies（频数） 8, 23, 67–68, 73
comparing（比较） 14
continuous data（连续数据） 328
cumulative frequency（累积频数） 34–38, 42
highest frequency group of values（具有最高频数的类） 52
histograms（直方图） 24–30
percentages with no frequencies（无频数百分数） 12
frequency density（频数密度） 27–32, 68
Frequency Density Up Close（频数密度细细看） 29
frequency scales（频数刻度） 13

G

Gaussian distribution（高斯分布） 352
geometric distribution（几何分布） 277–287, 297, 301, 324
guide（指南） 284
inequalities（不等式） 279
pattern of expectations（期望模式） 280–281
variance（方差） 281–284
Geometric Distribution Up Close（几何分布细细看） 278
goodness of fit（拟合优度） 573
test（检验） 579
graphs（图形，参见：charts and graphs）
grouped data（分组数据） 19

H

height probabilities（身高概率） 338–341
histograms（直方图） 19–28
frequency（频数） 24–30, 25
intervals（区间） 20
making（使得） 20
making area proportional to frequency（使面积与频数成比例） 26–28
mean（均值） 56
unequal intervals（不等宽区间） 24–30
when not to use（何时不用） 33
horizontal bar charts（水平条形图） 11, 23
horse racing（赛马） 243–246
hypothesis tests（假设检验） 521–566
alternate hypothesis（备择假设） 529–530, 543
critical region（拒绝域） 531–534, 539, 548
critical value（临界值） 532
null hypothesis（原假设） 528, 543
one-tailed tests（单尾检验） 534, 539
p-value（p值） 539
power of a hypothesis test（假设检验的功效） 561
process（过程） 526–539
overview（总览） 527
Step 1: Decide on the hypothesis（第1步：确定要进行检验的假设） 528–529, 543
Step 2: Choose the test statistic（第2步：选择检验统计量） 531, 544
Step 3: Determine the critical region（第3步：确定用于做决策的拒绝域） 532, 548
Step 4: Find the p-value（第4步：求检验统计量的p值） 535–536
Step 5: Is the sample result in the critical region?（第5步：查看样本结果是否处于拒绝域内） 537
Step 6: Make your decision（第6步：做出决策） 537
significance level（显著性水平） 533, 538, 539
statistically significant（统计显著性） 551
test statistic（检验统计量） 531, 539, 544, 547
two-tailed tests（双尾检验） 534, 539
Type I error（第一类错误） 555–560, 566
Type II error（第二类错误） 555–560, 566

I

incorrect sampling unit（抽样单位不正确） 425

independence（独立性） 573
independent events（独立事件） 182–183, 189–190
  versus mutually exclusive（互斥） 183
independent observations（独立观察结果） 224–226, 377, 472
  expectation（期望） 378
  of X（X的……） 233
  variance（方差） 378
  versus linear transforms（……与线性变换） 376–378
independent random variables（独立随机变量） 230–233, 368
indpendent variables（独立变量） 608, 646
information（信息）
  versus data（…… 与数据） 5
  visualizing（图形化，参见：visualizing information）
interpercentile range（百分位距） 98, 102
interquartile range（四分位距） 92–93, 97
  average distance（平均距离） 105
  versus the median（……与中位数） 97
intersecting events（相交事件） 147–154
intersection（交集） 149–154

K

kth percentile（第k百分位数） 99, 102

L

Law of Total Probability（全概率公式） 172, 178
least squares regression（最小二乘回归法） 626, 648
Least Squares Regression Up Close（最小二乘回归法细细看） 626
leaves（叶） 644
left-skewed data（左偏斜数据） 62, 64
letters, using to represent numbers（用字母表示数字） 48–49
linear correlations（线性相关） 613, 630–631
Linear Correlations Up Close（线性相关细细看） 613
linear regression（线性回归） 626, 640, 650
linear relationship between E(X) and E(Y)（E(X)与E(Y)之间的线性关系） 217–218
linear transforms（线性变换） 219, 220, 224–226
  distribution（分布） 376
  expectation and variance（期望与方差） 233
  versus independent observations（……与独立观察结果） 376–378
  versus playing multiple games（……与多玩几局赌博游戏） 221
line charts（线形图） 41, 42
Line Charts Up Close（线形图细细看） 41
line of best fit（最佳拟合线） 618, 622, 640
  finding equation（求公式） 622
  finding slope（求斜率） 623–624
  minimizing errors（误差最小化） 620–621
  non-linear（非线性） 650
  sum of squared errors（误差平方和） 620–621
lower bounds（下界） 86, 97
  basketball scores（篮球赛得分） 88
lower quartile（下四分位数） 92
  finding（求……） 94

M

matched pairs design (experiments)（配对设计） 647
mean（均值） 47–60
  basketball scores（篮球赛得分） 88
  binomial distribution（二项分布） 389
  calculating（计算） 50
  calculating when to use（计算何时使用） 78
mean (continued)（均值（续））
  categorical data（类别数据） 62
  distributions（分布） 56
  frequencies（频数） 52
  frequency density（频数密度） 68
  histograms（直方图） 56
  of two middle numbers（两个中间数） 61
  outliers（异常值） 57–59
  positive and negative distances（正负距离） 105
  problems with（问题） 65–72
  skewed data（偏斜数据） 62, 64
  standard deviations from（标准差） 121
  using letters to represent numbers（用字母表示数字） 48–49
  versus median（……与中位数） 62
  X + Y 368
  μ（缪） 50
  Σ（西格玛） 49

measuring probability（量度概率） 132
median（中位数） 61–70
calculating when to use（计算何时使用） 78
categorical data（类别数据） 62
frequency density（频数密度） 68
in three steps（三步法） 62
middle quartile（中间的四分位数） 92
problems with（问题） 65–72
skewed data（偏斜数据） 64
versus mean（均值） 62
versus the interquartile range（……与四分位距） 97
middle quartile（中间的四分位数） 92
modal class（众数组） 73
mode（众数） 73–80
calculating when to use（计算何时使用） 78
categorical data（类别数据） 73
three steps for finding（求…三步法） 74
mu（参见：μ（缪））
multiple sets of data（多批数据） 14, 23
mutually exclusive events（互斥事件） 147, 150

N

n! 248
negative linear correlation（负线性相关） 613, 631, 640
no correlation（不相关） 613, 631
No Dumb Questions（世上没有傻问题）
adding probabilities（概率相加） 143
alternate hypothesis（备择假设） 530
approximating binomial distribution（近似二项分布） 398
arranging objects in circle（对象环形排列） 248
average distance（平均距离）
interquartile range（四分位距） 105
Bayes' Theorem（贝叶斯定理） 179
bias（偏倚） 426, 434
binomial distribution（二项分布） 301, 412
bivariate data（二变量数据） 616
box and whisker diagram（箱线图） 101
breaking data into more than four pieces（将数据分割为四块以上） 97
central limit theorem（中心极限定理） 485
charts（图表） 5
clustered sampling（整群抽样） 434
confidence intervals（置信区间） 491, 518, 539
confidence interval versus confidence level（置信区间与置信水平） 507
continuity corrections（连续性修正） 398, 412
continuous data（连续数据） 370
continuous distributions（连续分布） 352
correlation coefficient（相关系数） 634
cumulative frequency（累积频数） 36
degrees of freedom（自由度） 576, 595
discrete data（离散数据） 370
discrete random variable（离散随机数据） 203
distribution of X + Y（X+Y的分布） 370
drawing lots（抽签） 434
$E(X_1 + X_2)$ and E(2X)（E(X1 + X2)与E(2X)） 224
expectation（期望） 208, 219
factorials（阶乘） 248
frequency density（频数密度） 30
Gaussian distribution（高斯分布） 352
geometric distribution（几何分布） 277, 284, 301
histograms（直方图） 23, 30
how data is spread out（数据分散方式） 97
hypothesis tests（假设检验） 530, 552
independent events（独立事件） 184
independent observations（独立观察结果） 378
independent versus mutually exclusive（独立与互斥） 184
information versus data（信息与数据） 5
interquartile range（四分位距） 97
limit on intersecting events（相交事件） 154
linear transforms（线性变换） 219, 378
line charts（线形图） 42
line of best fit（最佳拟合线） 624
mean or median with categorical data（类别数据的均值或中位数） 62
mean with skewed data（有偏斜数据的均值） 62
median（中位数） 352
versus mean（…… 与均值） 62
versus the interquartile range（…… 与四分位距） 97
n! 248
normal distribution（正态分布）
accuracy of（…… 的精确性） 398
approximating binomial or Poisson distribution（近似二项分布或泊松分布） 412
normal probability tables（正态概率表） 352
null hypothesis（原假设） 530
outliers（异常值） 634
P(Black | Even)（P(黑 | 偶)） 179
permutations and combinations（排列与组合） 263

arranging by type（按种类排列） 257
point estimators（点估计量） 446, 452
and sampling distributions（与抽样分布） 485
Poisson distributions（泊松分布） 311, 314, 412
approximating binomial distribution（近似二项分布） 317, 398
population mean（总体均值） 446
positive and negative distances（正负距离） 105
probabilities written as fractions, decimals, or percentages（以分数、小数表示概率或百分数） 139
probability（概率） 139
best method（最佳方法） 143
probability density function（概率密度函数） 334
probability distributions（概率分布） 203
letters p and q（字母p和q） 284
quiz show（智力游戏节目） 290
probability for standardized range（标准化数值范围的概率） 347
probability of range（数值范围概率） 352
probability tables（概率表） 352, 370
probability trees（概率树） 165, 179
proportion versus probability（比例与概率） 456
questionnaires（调查问卷） 426
random variables（随机变量） 233
right– and left–skewed data（左右偏斜数据） 62
roulette wheel（轮盘赌） 184
sample mean（样本均值） 446
sample variance（样本方差） 452
sampling bias（抽样偏倚） 434
sampling distribution（抽样分布） 466
sampling frame（抽样框架） 426
scatter diagrams（散点图） 616
set theory（集合论） 139
shortcuts（简捷算法） 370
significance level（显著性水平） 539
significance tests（显著性检验） 552
slot machines（老虎机） 208
standard deviation（标准差） 113, 122, 208
standard error（标准误差） 485
of proportion（比例） 466
standard scores（标准分） 122, 347, 352
outliers（异常值） 122
statistical sampling（统计抽样）
bias（偏倚） 426
clustered sampling（整群抽样） 434
drawing lots（抽签） 434
increasing sample size（增大样本） 434
simple random sampling（简单随机抽样） 434
stratified sampling（分层抽样） 434
stratified sampling（分层抽样） 434
systematic sampling（系统抽样） 434
t–distributions（t分布） 518
target population（目标总体） 426
Type I error（第一类错误） 560
Type II error（第二类错误） 560
variance（方差） 122, 208
variance equations（方差公式） 113
variances（方差） 219
Venn diagrams（维恩图） 139, 165, 184
χ^2 (chi square) distribution（χ^2（卡方）分布） 595
χ^2 (chi square) tests（χ^2（卡方）分布） 576
no linear correlation（非线形相关） 630
non–linear relationships（非线性关系） 650
normal approximation（正态近似） 394
normal distribution（正态分布） 325–360, 361–414
accuracy of（精确性） 398
approximating continuity correction（近似连续性修正） 396
approximating binomial distribution（近似二项分布） 386
approximating binomial or Poisson distribution（近似二项分布或泊松分布） 412
approximating binomial probabilities（近似二项概率） 397
binomial distribution（二项分布） 384, 389, 392–393
approximating（近似） 398, 407
continuous（连续） 395
continuous data（连续数据） 337, 365
continuous distributions（连续分布） 352
continuous probability distributions（连续概率分布） 337
defined（定义） 339–340
discrete data（离散数据） 337
discrete data versus continuous data（离散数据与连续数据） 326–327
empirical rule（经验法则） 645
finding ≤ probabilities（≤型概率的求解） 397
finding ≥ probabilities（≥型概率的求解） 397
finding between probabilities（“介于”型概率的求解） 397
frequency and continuous data（频数与连续数据） 328
Gaussian distribution（高斯分布） 352
height probabilities（身高概率） 338–341

in place of binomial distribution（代替二项分布） 389
median（中位数） 352
normal probability tables（正态分布表） 352
Poisson distribution（泊松分布） 386, 406
Pool Puzzle（奇妙池） 399–400
probability = area（概率 ＝ 面积） 331
probability density function（概率密度函数） 330–337, 337
probability for standardized range（标准化数值范围的概率） 347
probability of range（标准化数值范围） 352
probability tables（概率表） 349–352
standard score（标准分） 345–347, 352
table（表格） 411
transforming（变换） 345
versus binomial distribution（…… 与二项分布） 393, 395
versus t-distributions（…… 与t分布） 515
Normal Distribution Exposed（正态分布访谈） 404
normal probabilities（正态概率） 359
calculating（计算） 341–352
determining distribution（确定分布） 343
standardizing normal variables（正态变量标准化） 344
tables（表格） 349–352, 352, 658–659
nu（参见 ν （正态变量））
null hypothesis（原假设） 528, 530, 543
numbers, using letters to represent（数字，用字母表示） 48–49
numerical data（数字数据） 18, 23

O

observations（观察） 222–224
independent（独立） 224
shortcuts（速算法） 223
one-tailed tests（单尾检验） 534, 539
outliers（异常值） 57–59, 89–91, 93, 634
interquartile range（四分位距） 93
standard scores（标准分） 122

P

p-value（p值） 535–536, 539
percentage sales（百分数刻度） 12
percentages with no frequencies（无频数百分数） 12
percentiles（百分位数） 98–99, 102
kth percentile（第k百分位数） 99, 102
perfect negative linear correlation（完全负线性相关） 631
perfect positive linear correlation（完全正线性相关） 631
permutations and combinations（排列与组合） 241–268
arrangements（排位） 246
arranging by type（按种类排列） 252–257
arranging duplicates（重复排列） 254
arranging objects in circle（圆形排位） 247–248
combinations（组合） 260–263, 293
examining combinations（何为组合） 260–263
examining permutations（何为排列） 258–259
factorial（阶乘） 246
formulas for arrangements（排位方式的计算公式） 248
permutations versus combinations（排列与组合比较） 261
three-horse race（三马赛） 243–246
pie charts（饼图） 8–9, 9, 23
placebo（安慰剂） 646
point estimators（点估计量） 443–447, 452, 493, 519
and sampling distributions（抽样分布） 485
for population variance（总体方差） 457
problem with（问题） 489
Poisson distribution（泊松分布） 306–319, 324, 386, 406, 407, 412
approximating binomial distribution（近似二项分布） 398
approximating the binomial distribution（近似二项分布） 316–317
central limit theorem（中心极限定理） 482
expectation and variance（期望与方差） 308
guide（指南） 319
when λ is large（当λ很大） 407
when λ is small（当λ很小） 407
X + Y 312–313
Poisson Distribution Up Close（泊松分布细细看） 307
Poisson variables, combining（泊松变量，组合） 313
Pool Puzzle（奇妙池）
binomial distribution（二项分布） 299–300
confidence intervals（置信区间） 499–500
continuity correction（连续性修正） 399–400
discrete probability distributions（离散概率分布） 215–216

population（总体） 418, 438
chart（图表） 419
mean（均值） 446
proportion（比例） 454–455, 457
variance（方差） 448–450
versus samples（…… 与样本） 418
（同时参见总体和样本的估计）
positive and negative distances（正负距离） 105
positive linear correlation（正线性相关） 613, 631, 640
possibility space（概率空间） 135
precision, problem with（精度，问题） 489
probability（概率） 127–196
= area（等于面积） 331
adding（相加） 142, 143
Bayes' Theorem（贝叶斯定理） 173, 178
best method（最佳方法） 143
conditional（条件） 157–160
probability tree（概率树） 158–161
events（事件，参见：events）
for a sample（用于样本） 459
how probability relates to roulette（概率与轮盘赌的关系） 132
intersection（交集） 149–154, 153
Law of Total Probability（全概率公式） 172, 178
measuring（量度） 132
of getting a black or even（出现黑色或偶数……） 145–146
proportion（比例） 455
range of values（数值范围） 329
union（并集） 149–154, 153
Venn diagram（维恩图） 136, 154
written as fractions, decimals, or percentages（记作分数、小数或百分数） 139
probability density（概率密度） 334
function（函数） 330–337
never equaling 0（永远不会等于0） 341
probability distributions（概率分布） 220, 224, 363
4X 376
binomial（二项，参见：binomial distribution）
continuous data（连续数据） 329–333
geometric（几何，参见：geometric distribution）
large number of possibilities（大量概率） 273, 277
letters p and q（字母p和q） 284
new price and payouts（新价码与赔率） 212–214
normal（正态，参见：normal distribution）
of X + Y（X+Y） 372
patterns（固定模式） 274–277
Poisson (see Poisson distribution)（泊松）
random variable X（随机变量X） 210
standard deviation（标准差） 207
Probability Distributions Up Close（概率分布细细看） 202
probability tables（概率表） 349–352, 352, 370, 513, 657–661
standard normal probabilities（标准正态概率） 658–659
t-distribution critical values（t分布临界值） 660
χ^2 (chi square) critical values（χ^2（卡方）临界值） 661
Probability Tables Up Close（概率表细细看） 351
probability trees（概率树） 158–161, 165, 180
hints（决策） 161
proportions（比例） 9
probability（概率） 455
sampling distribution of（抽样分布） 460
distribution of P_s（P_s分布） 464–466
expectation of P_s（P_s期望） 462
variance of P_s（P_s方差） 463
standard error of（标准值） 463

Q

qualitative data（定性数据） 18
quartiles（四分位数） 92
interquartile range（四分位距） 92–93
lower（下） 92, 94
middle（中） 92
upper（上） 92, 94
questionnaires, bias（调查问卷，偏倚） 426

R

randomization（随机化） 646
randomized block design (experiments)（随机化区组设计） 647
random number generators（随机编号生成器） 431
random variables（随机变量） 202
adding（加） 230
continuous（连续） 331
independent（独立） 233
subtracting（减） 231

range（…距） 86–103, 97, 329, 333
basketball scores（篮球赛得分） 88
calculating（计算） 86
lower bound（下界） 86
outliers（异常值） 89–91
problems with（问题） 90
quartiles（四分位数） 92
upper bound（上界） 86
regression（回归，参见：correlation and regression）
replication（复制） 646
response variable（反应变量） 608
right-skewed data（右偏斜数据） 62, 64
roulette（轮盘赌） 129–196
black and even pockets（黑色和偶数球位） 156
board（轮盘板） 129–130
how probability relates to（概率与…的关系） 132
independent events（独立事件） 184
measuring probability（量度概率） 132
P(Black | Even)（P(黑|偶)） 167–171
P(Even)（P(偶)） 169
possibility space（概率空间） 135
probabilities（概率） 135
probability of ball landing on（停球结果为7的概率） 7 133–134
sample space（样本空间） 135

S

samples（样本） 418, 438
biased（偏倚） 424–426
designing（设计） 422–423
mean（均值） 445, 446
space（空间） 135
survey（调查） 418, 438
unbiased（无偏倚） 424–426
unreliability（不可靠） 420
variance（方差） 449, 452
（同时参见估计总体与样本）
sampling（抽样，参见：statistical sampling）
sampling distribution（抽样分布） 466
difference between two means（两个均值之间的差异） 652
difference between two proportions（两个比例之间的差异） 653
sampling distribution of means（均值的抽样分布） 471–479
distribution of x（X的分布） 480
variance of X（X的方差） 476
sampling distribution of proportion（比例的抽样分布） 460
distribution of P_s（P_s的分布） 464–466
expectation of P_s（P_s的期望） 462
variance of P_s（P_s的方差） 463
Sampling Distribution of Proportions Up Close（比例的抽样分布细细看） 469
Sampling Distribution of the Means Up Close（均值的抽样分布细细看） 479
sampling frame（抽样框架） 423–428, 438
bias（偏倚） 425
sampling units（抽样单位） 422, 428
bias（偏倚） 425
sampling without replacement（不重复抽样） 430
sampling with replacement（重复抽样） 430
scales（刻度） 12
scatter diagrams（散点图） 609, 612, 616, 618, 640
line of best fit（最佳拟合线） 618
finding equation（求方程） 622
finding slope（求斜率） 623–624
sum of squared errors（误差平方和） 620–621
scatter plots (see scatter diagrams)（散点图）
segmented bar chart（分段条形图） 14
set theory（集合论） 139
shortcuts（简捷算法） 370
sigma (Σ)（西格玛（Σ）） 49
sigma (σ)（西格玛（σ）） 107
significance level（显著性水平） 533, 538, 539
significance tests（显著性检验） 552
simple random sampling（简单随机抽样） 430–431, 434, 436, 438
drawing lots（抽签） 431
random number generators（随机编号生成器） 431
skewed data（偏斜数据） 58–59, 64
mean（均值） 62
Skewed Data Up Close（偏斜数据细细看） 59
skewed to the left（左偏斜） 59
skewed to the right（右偏斜） 58–59
slope of regression line（回归线斜率）

confidence intervals（置信区间） 651
slot machines（老虎机） 198
discrete random variables（离散随机变量） 202
low versus high variance（低方差与高方差） 208
probability distributions（概率分布） 201
variance（方差） 207
split-category bar chart（分立条形图） 14
standard deviation（标准差） 107–110, 113–117, 207, 220
from the mean（从均值） 121
variance equations（方差公式） 113
σ (sigma)（σ（西格玛）） 107, 224
Standard Deviation Exposed（标准差访谈） 108
standard error（标准误差） 485
of mean（均值…） 479
of proportion（比例…） 463, 466
standardizing normal variables（正态变量标准化） 344
standard normal probabilities（标准正态概率） 658–659
standard scores（标准分） 118–122, 345–347, 352
calculating（计算） 119
interpreting（解释） 120
Standard Scores Up Close（标准分细细看） 121
statistical sampling（统计抽样） 415–440
bias in sampling（抽样偏倚） 423–426, 434, 438
sources（来源） 425
choosing samples（选择抽样） 430
cluster sampling（整群抽样） 433, 433–434, 436, 438
defined（确定） 418
designing samples（设计样本） 422
drawing lots（抽签） 431, 434
how it works（抽样方法） 419
incorrect sampling unit（抽样单位不正确） 425
increasing sample size（增大样本） 434
population（总体） 418, 438
population chart（总体图） 419
populations versus samples（总体与样本） 418
random number generators（随机编号发生器） 431
representative sample（代表性样本） 420
samples（样本） 438
unreliability（不可靠） 420
sample survey（样本调查） 418, 438
sampling bias（抽样偏倚） 434
sampling chart（抽样图） 419
sampling frame（抽样框架） 423–428, 438
sampling units（样本单位） 422, 428
sampling without replacement（不重复抽样） 430
sampling with replacement（重复抽样） 430
simple random sampling（简单随机抽样） 430–438
choosing（选择抽样） 431
strata（层） 432
stratified sampling（分群抽样） 432, 434, 436, 438
systematic sampling（系统抽样） 433–434, 438
target population（目标总体） 422, 428, 438
unreliability（不可靠） 420
statistics（统计量）
defined（定义） 2
why learn（为何学习） 3
statistics tables（统计表） 657–661
standard normal probabilities（标准正态概率） 658–659
t-distribution critical values（t分布临界值） 660
χ^2 (chi square) critical values（χ^2(卡方)临界值） 661
stemplots（茎叶图） 644
stems（茎） 644
strata（层） 432
stratified sampling（分群抽样） 432–438
stratified sampling（分群抽样） 436
summation symbol (Σ)（求和符号（Σ）） 49
sum of squared errors（误差平方和） 640
symmetric data（对称数据） 59
systematic sampling（系统抽样） 433–434, 438

T

t-distributions（t分布） 509–515
probability tables（概率表） 513
shortcuts（简便方法） 515
small sample（小样本） 510
standard score（标准分） 511
table（表） 660
versus normal distributions（……与正态分布） 515
target population（目标总体） 422, 426, 428, 438
test statistic（检验统计量） 531, 539, 544, 547
three-horse race（三马赛） 243–246
two-tailed tests（双尾检测） 534, 539
Type I error（第一类错误） 555–560, 566
Type II error（第二类错误） 555–560, 566

U

unbiased sample（非偏倚样本） 424–425

uniform distribution（均匀分布） 655

union（并集） 149–154

univariate data（单变量数据） 608, 640

upper bounds（上界） 86, 97

  basketball scores（篮球赛得分） 88

upper quartile（上四分位数） 92

  finding（求解） 94

V

variability（差异性） 104–124

  average distance（平均距离） 105

  positive and negative distances（正负距离） 105

  variance（方差，参见：variance）

variables（变量） 368

  probabilities involving the difference between two（两个变量之差的概率） 369

variance（方差） 106–113, 122, 205–208, 219, 220, 367

  binomial distribution（二项分布） 298, 389

  calculating（计算） 111–113

    quicker way（更快方法） 113

  geometric distribution（几何分布） 281–284

  independent observations（独立观察结果） 378

  linear transforms（线性变换） 233

  of X（X……） 476

  Poisson distribution（泊松分布） 308

  slot machines（老虎机） 207

  standard deviation（标准差） 107–110

    σ (sigma)（σ（西格玛）） 107

  two games（两局赌局） 222–224

  X + Y 368

Variance Up Close（方差细细看） 450

Venn diagrams（维恩图） 136, 139, 154, 165

  conditional probability（条件概率） 157

  independent events（独立事件） 184

vertical bar charts（垂直条形图） 10–11, 23

visualizing information（信息图形化） 1–44, 19–28

  categorical and numerical data（类别数据与数字数据） 18–23

  cumulative frequency（累积频数） 34–38

  histograms（直方图） 19–28

  statistics（统计量） 2

  （参见图形图表）

Vital Statistics（重要统计量）

  A or B（A或B） 153

  approximating binomial distribution（近似二项分布） 389

  approximating Poisson distribution（近似泊松分布） 407

  arranging by type（按种类排列） 254

  Bayes' Theorem（贝叶斯定理） 178

  combinations（组合） 263

  conditions（条件概率） 165

  cumulative frequency（累积频数） 34

  event（事件） 132

  formulas for arrangements（排位方式的计算公式） 248

  frequency（频数） 8

  independence（独立） 184

  independent observations（独立观察结果） 224

  interquartile range（四分位距） 93

  Law of Total Probability（全概率公式） 178

  linear transforms（线性变换） 220

  mean（均值） 54

  mode（模式） 76

  outlier（异常值） 58

  percentile（百分位数） 99

  permutations（排列） 263

  probability（概率） 143

  quartiles（四分位数） 92

  range（距） 86

  significance level（显著性水平） 533

  skewed data（偏斜数据） 58

  standard score（标准分） 346

  uniform distribution（均匀分布） 655

  variance（方差） 106, 113

W

Watch it!（小心！）

  criteria of np>10 and nq >10（条件：np>10与nq >10） 389

  cumulative frequencies（累积频数） 35

  exclusive versus exhaustive（互斥与穷举） 150

  how large n needs to be（n需要有多大） 465

  independent random variables（独立随机变量） 230–232

  independent versus mutually exclusive（独立与互斥） 183

  linear regression（线性回归） 626

percentages with no frequencies（无频数百分数） 12
quartiles（四分位数） 92
samples equation（样本公式） 451
subtracting random variables（减去随机变量） 231
$X_1 + X_2$ and 2X（$X_1 + X_2$ 和 2X） 223
Who Wants To Win A Swivel Chair（转椅赢赢赢） 289, 381–386
expectation and variance（期望与方差） 304
generalizing probability for three questions（推而广之至求3个问题的概率） 293
generalizing the probability（进一步推导概率算式） 296
probability of getting exactly three questions right（答对三题的概率） 304
probability of getting exactly two questions right（答对两题的概率） 304
probability of getting no questions right（一题也答不对的概率） 304
probability of getting two or three questions right（答对两题或三题的概率） 304
should you play or walk away（玩下去，还是转身走） 291
width of data（数据宽度） 88

X

X + Y Distribution Up Close（X+Y分布细细看） 368
X – Y Distribution Up Close（X – Y分布细细看） 369

Z

z–scores（z分） 118–122
calculating（计算） 119
interpreting（释义） 120

反侵权盗版声明

电子工业出版社依法对本作品享有专有出版权。任何未经权利人书面许可，复制、销售或通过信息网络传播本作品的行为，歪曲、篡改、剽窃本作品的行为，均违反《中华人民共和国著作权法》，其行为人应承担相应的民事责任和行政责任，构成犯罪的，将被依法追究刑事责任。

为了维护市场秩序，保护权利人的合法权益，我社将依法查处和打击侵权盗版的单位和个人。欢迎社会各界人士积极举报侵权盗版行为，本社将奖励举报有功人员，并保证举报人的信息不被泄露。

举报电话：（010）88254396；（010）88258888

传　　真：（010）88254397

E－mail：dbqq@phei.com.cn

通信地址：北京市万寿路173信箱

电子工业出版社总编办公室

邮　　编：100036

[General Information]
书名=深入浅出统计学
作者=（美）格里菲思著
页数=679
SS号=12963059
出版日期=2012.01
出版社=北京市：电子工业出版社

封面
书名
版权
前言
目录
1 信息图形化：第一印象
2 集中趋势的量度：中庸之道
3 分散性与变异性的量度：强大的“距”
4 概率计算：把握机会
5 离散概率分布的运用：善用期望
6 排列与组合：排序、排位、排
7 几何分布、二项分布及泊松分布：坚持离散
8 正态分布的运用：保持正态
9 再谈正态分布的运用：超越正态
10 统计抽样的运用：抽取样本
11 总体和样本的估计：进行预测
12 置信区间的构建：自信地猜测
13 假设检验的运用：研究证据
14 X2分布：继续探讨……
15 相关与回归：我的线条如何？
附录I 尾声：正文未及的十大拓展
附录II 统计表：快来查表